PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-051806
(43)Date of publication of application : 23.02.2001

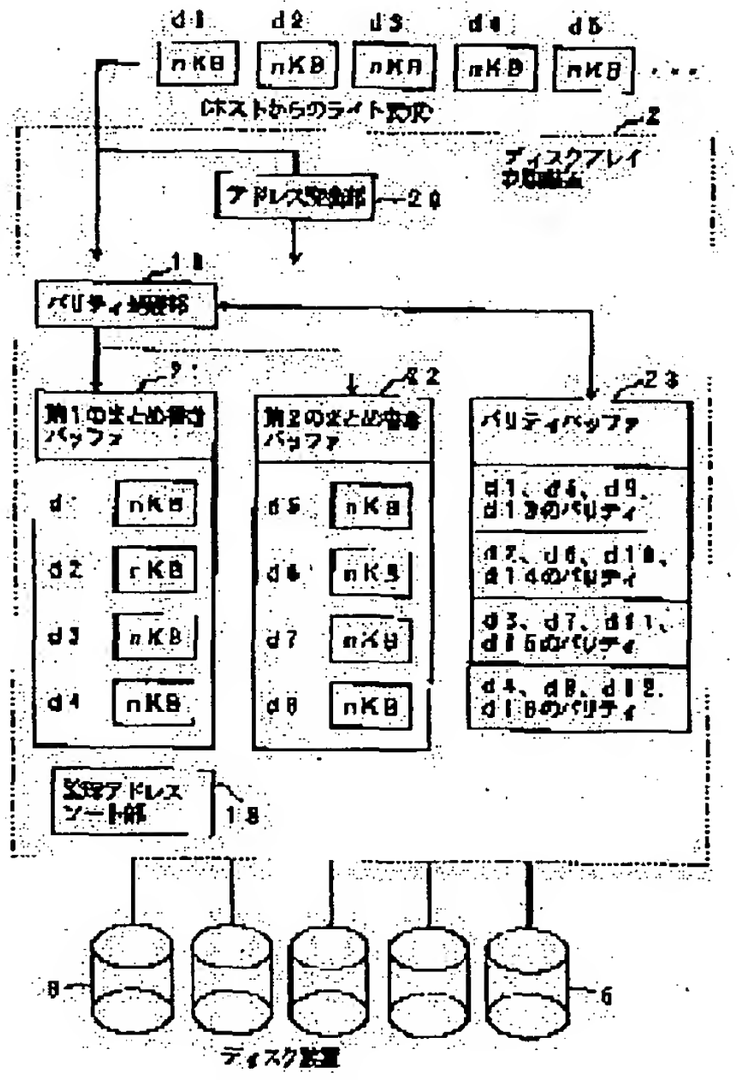
(51)Int.Cl. G06F 3/06

(21)Application number : 11-222078
(22)Date of filing : 05.08.1999

(71)Applicant : FUJITSU LTD
(72)Inventor : OTA YOSHIYUKI
NISHIKAWA KATSUHIKO
AOKI TAKAHIRO

(54) DISK ARRAY DEVICE

(57)Abstract:
PROBLEM TO BE SOLVED: To improve the overhead and read performance of a disk array RAID (Redandant Array of Inexpensive Disks) at write time and to efficiently perform the generation management of files.
SOLUTION: This device is equipped with a parity process part 19 which finds parity data by exclusively ORing data of a write request from a host, 1st and 2nd batch write buffers 21 and 22 which temporarily hold the data of the write request from the host and have double buffer constitution, a parity buffer which holds the parity data until all data constituting stripes are written to the batch write buffers, an address conversion table which holds the correspondence between logical addresses and physical addresses, and a logical address sorting part 18 which sorts the data in the batch write data by using logical keys as keys.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

* NOTICES *

Japan Patent Office is not responsible for any
damages caused by the use of this translation.

1. This document has been translated by computer.So the translation may not reflect the original precisely.
2. **** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1] Disk array equipment which is equipped with the disk array control unit which is characterized by providing the following, and which performs writing / read-out control of data between two or more disk units and each disk unit, two or more disk units are made to distribute the division data of a block unit, stores, and stores in one of disk units the parity data for which it asked from these division data with this disk array control unit. The parity processing section which computes the exclusive OR of the data of the light demand from a host, and asks for parity data. The 1st, the 2nd conclusion writing buffer which hold the data of the light demand from a host temporarily, and take double buffer composition. The parity buffer which holds parity data until all the data that constitute a stripe are written in in the aforementioned conclusion writing buffer. The logical address sorting section which carries out the aforementioned logical address at a key, and sorts the address translation table holding the correspondence information on the logical address specified by the host and the physical address in the disk with which data are actually stored, and the data in the aforementioned conclusion writing buffer.
[Claim 2] Disk array equipment according to claim 1 characterized by what is characterized by providing the following. The 1st control means which are not concerned with new and updating but store the light data from a host in the conclusion writing buffer of one of the above in order of arrival. When a lead demand occurs to the same disk as the disk which stores the data in the aforementioned conclusion writing buffer after the time when this conclusion writing buffer became full, If it judges whether the free area which stores all the data in the aforementioned conclusion writing buffer is near the position on the disk with which the corresponding lead data are stored and there is the aforementioned free area, he has no seeking in the free area. The 2nd control means which became the aforementioned full and which collect and write and store the light data of a buffer.
[Claim 3] Disk array equipment which is equipped with two or more disk units characterized by providing the following, and the disk array control unit which performs writing / read-out control of data in operating each disk unit in parallel, two or more disk units are made to distribute the division data of a block unit, stores, and stores in one of disk units the parity data for which it asked from these division data with this disk array control unit. The parity processing section which was led when an updating light demand occurred from a host and which computes the parity after updating from the data before updating and parity data in a corresponding disk. The front [updating] data buffer which stores the data before updating led from the disk with the address value in the disk. The generation-control section which stores the data in this buffer in a disk when the data after updating in the same position on a disk and the parity after updating are stored when disk media take at least 1 round and the data buffer before updating becomes full, after leading the data before updating, and parity from a disk.
[Claim 4] When an updating light demand occurs from a host, the aforementioned generation-control section leads the data before updating and parity from a disk, and makes the parity after updating in the aforementioned parity processing section compute. When disk media take at least 1 round after leading the data before updating, and parity from a disk The data before updating which stored the data and parity after updating in the same position on a disk, and were led from

the disk Disk array equipment according to claim 3 characterized by what it has for the control means which store in the aforementioned data buffer before updating with the address value in the disk, and store the data in this buffer in the continuation field in a disk when this data buffer before updating becomes full.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[The technical field to which invention belongs] this invention relates to the disk array equipment of the RAID level 4 equipped with the disk array control unit which two or more disk units and aforementioned disk units are operated in parallel, and performs read-out / write-in control of data, or the RAID level 5.
[0002]
[Description of the Prior Art] Hereafter, the conventional example is explained.
[0003] **1: Explanatory-drawing 6 reference drawing 6 of disk array equipment is explanatory drawing of conventional disk array equipment. Disk array equipment is carrying out parallel operation of two or more built-in magnetic disk units (hard disk drive unit), and is the external storage which attained improvement in the speed of read-out/drawing speed of data, and raised reliability by introduction of a redundant configuration, or auxiliary memory. In addition, the following explanation only describes a "disk unit" the aforementioned magnetic disk unit (or hard disk drive unit).
[0004] It is the disk unit 6-1 from which disk array equipment constitutes the disk array control unit 2 and two or more RAID (it mentions later for details) as shown in drawing 6 , 6-2, and 6-3... It consists of 6-m and 6-n (n=m +1). Moreover, a host adaptor 3, the disk array controller 4 and two or more device (adapter DA) 5-1, 5-2, 5-3 ... 5-m and 5-n are prepared. [control unit / disk array / 2] And each disk unit 6-1 - 6-n are connected to the device adapter 5-1 - 5-n, respectively.
[0005] Although the aforementioned disk array equipment is connected and applied to a host 1, it connects between the host adaptors 3 of the disk array control unit 2 by the interface cable (for example, cable for SCSI-2) with a host 1 in this case. A host adaptor 3 performs the interface control to a host 1, and the disk array controller 4 performs various control at the time of the read/write of data etc. Device adapter (DA) 5-1 - 5-n perform control to a disk unit 6-1 - 6-n with directions of the disk array controller 4 at the time of the read/write of data.
[0006] When disk array equipment is seen from a host 1, it is visible to one set of a disk unit. With this disk array equipment, if the data with which the host adaptor 3 was sent by the host 1 are received for example, the data will be sent to the disk array controller 4. And the disk array controller 4 creates the parity data to the aforementioned data, and stores them in the remaining one-set of disk unit 6-n (disk unit for parity) through device adapter 5-n while it divides for example, the aforementioned data into two or more data and stores them in two or more disk units 6-1 - 6-m (disk unit for data) through the device adapter 5-1 - 5-m.
[0007] Thus, disk array equipment can realize improvement in the speed of read/write from one set of a disk unit by writing the data of big size in two or more disk units simultaneously, or reading from two or more disk units simultaneously, and can raise the reliability of data. Therefore, highly efficient-ization of equipment can be attained.
[0008] **2: The explanation aforementioned disk array equipment of RAID is a method which realizes reliability and a performance higher than an independent disk unit by using two or more disk units (hard disk drive unit). This will be David .A. Paterson (David.A.Patterson) of University of California at Berkeley of the U.S. in 1987. It is referred to as RAID (Redundant Arrays of Inexpensive Disks) which professors advocated. That is, Above RAID is how originating in the paper of aforementioned Professor David .A. Paterson to call.
[0009] As mentioned above, the disk array method RAID which accesses a lot of [at high speed] data at many disk units, and realizes the redundancy of the data at the time of disk failure It is classified into the level (the aforementioned level is hereafter described also as RAID1, RAID2, RAID3, RAID4, and RAID5) from one to five. On level 5 (RAID3, RAID4, RAID5), the parity data for recovering data at the time of failure of a disk unit are held from level 3.
[0010] Also in the level of Above RAID, further, in level 5 (RAID5), two or more simultaneous read-out is possible, it is not fixing the disk unit which stores parity, and two or more simultaneous writing is also enabled and an effect is demonstrated in a lot of transaction processing on level 4 (RAID4) and level 5 (RAID5).
[0011] **3: Explanatory-drawing 7 reference drawing 7 of the level of RAID is level explanatory drawing of RAID, and ** view is drawing having ** shown [the data transfer (writing) of RAID 3 and 4, and] the data transfer (writing) of RAID5. In addition, for 3, as for a disk array controller, and A, B, C, D and E, in drawing 7 , a host adaptor and 4 are [a device adapter (DA) and 6-1 to 6-5] disk units, respectively.
[0012] (1) : the explanation RAID 1 of RAID1 is the so-called MIRADO disk composition which doubled the disk unit. That is, the same data as two disk units are written. Although the cost of a disk unit cuts in double precision, it is the simpleest and there is also an actual result. Although the execution time is prolonged for a while about a performance in order to wait for the write-in completion to two disk units at the time of the writing of data, since a principle top is an execute permission if one of two disk units is vacant at the time of reading of data, it becomes the improvement in a performance.
[0013] (2) : the explanation RAID 2 of RAID2 divides input data (striping), and it divides and stores in two or more disk units, applying it. In this case, the disk unit which stores the error correction sign to the data which carried out [aforementioned] division is made into a redundant disk unit, and let the data of the aforementioned redundant disk unit be Hamming code.
[0014] When the number of the disk units for data storage is four, for example, three disk units for error correction signs are required of RAID2. It is that data are not lost with two or more disk unit obstacles as a feature of a disk unit, either. Therefore, as mentioned above, two or more redundant disk units are needed, and there is a fault that the net amount of data decreases.
[0015] (3) : ** view reference RAID 3 of the explanatory drawing 7 of RAID3 distributes and stores in two or more disk units the data which divided the input data (striping), and applied to which and divided the interleave. In this case, the disk unit which stores the error correction sign to the data which carried out [aforementioned] division is made into a redundant disk unit, and let the data of the aforementioned redundant disk unit be parity data.
[0016] For this reason, it is not concerned with the number of the disk units for data, but the number of the aforementioned redundant disk unit (disk unit for parity data) can be managed with RAID3 at one set. For example, in RAID3, as shown in ** view of drawing 7 , read/write of the data is carried out in parallel to two or more disk units 6-1 to 6-4, and it is characterized by having one set of the disk unit 6-5 for parity.
[0017] A data transfer rate becomes N times (number of the disk unit for data which carries out N:parallel operation) with the degree of parallel. Moreover, even if there is no degradation even if an obstacle occurs in one set of a disk unit, and an obstacle occurs in data transfer theoretically, operating as it is is also possible.
[0018] (4) : ** view reference RAID 4 of the explanatory drawing 7 of RAID4 makes the striping unit (division unit) in RAID3 a sector unit (1 or number sector). having made it that is, more advantageous on a performance for the disk unit which is each to distribute access in access of the small amount of data, although the effect on a performance is seen in the above RAID 3 when accessing in a to some extent big unit, in order to bundle two or more disk units and to access them
[0019] Although data are written independently of each disk unit by RAID4 from this, it has a

parity disk unit and the parity generated from the bit to which each disk unit corresponds is stored. Therefore, data are not lost with the obstacle of one set of a disk unit, either. A fault is having to access both data disk equipment and a parity disk unit at the time of renewal of data, and the parity disk unit in this case tends to grow into a bottleneck.
[0020] (5) : ** view reference RAID 5 of the explanatory drawing 7 of RAID5 makes each disk unit 6-1 to 6-5 distribute parity data, in order to cancel the access concentration to the parity disk unit which was the fault of RAID4. However, at the time of renewal of data, two disk array equipments must still be accessed, and on a principle, since parity data are ungenerable if they do not take difference with the old data, they need to take the process of a lead (read-out), data generation, and a light (writing).
[0021] Furthermore, although it can say that data loss was lost as a matter of fact with parity data, when one set of disk array equipment breaks down, the original equipment performance cannot be maintained at least. It is because the data of the disk array equipment which remained in making the data of broken disk array equipment are mobilized fully. Therefore, the thing which does not want to stop a system with the obstacle of disk array equipment is not turned to like a nonstop system.
[0022] One example of the data transfer (writing) of the above RAID 5 is as having been shown in ** view of drawing 7 . **. In a view, P0, P1, P2, and P3 are parity, and they are changing the disk unit which stores parity for every sector group of parity generation. In addition, a striping unit is a sector unit (1 or number sector) like [RAID5] RAID4.
[0023] (6) : there are some which divide data per block, and each disk which constitutes an array is made to distribute the divided data, and are memorized in the access method of the disk array equipment specified as the other explanation above RAID4 and RAID5. In this case, in the case of failure of one disk, in order to restore data, the exclusive OR of the data which were made to distribute and were memorized was calculated, and it has memorized on another disk as parity data.
[0024] When performing the writing for renewal of data to the disk array of this RAID 4 and 5, while reading the corresponding old data and the parity data corresponding to it at once and writing in updating data, it is necessary to write in by calculating new parity data. For this reason, compared with the data writing to the usual simple substance disk, excessive disk accessing is needed (light penalty).
[0025] **4: Explanatory-drawing 8 reference drawing 8 of LFS and WAFL is explanatory drawing of WAFL. As mentioned above, in a RAID array, although improvement in the throughput of external storage is aimed at by the parallel access, it writes in by existence of the above-mentioned light penalty (the aforementioned excessive disk accessing), and there is a problem that the overhead at the time will become large RAID4 and RAID5. On the other hand, there is technology called LFS (Log-Structured File System) or WAFL (Write Anywhere File Layout).
[0026] Above LFS and WAFL -- the data d1, d2, d3, d4, d5, and d6 of each light demand, for example, the light data of a nKB unit, ... is not concerned with new and updating, but it once stores in the conclusion writing buffer 10 in a disk controller, and a response is returned to a host
[0027] Then, when it collected and writes and a buffer 10 becomes full (full), it stores in the continuation field in a disk collectively. In this case, parity data are the light data d1, d2, d3, d4, d5, and d6 of the nKB unit which collects and writes and is stored in the buffer 10 one after another... It creates in between.
[0028] Moreover, in the address translation table 9, correspondence with the physical address by which data are actually remembered to be the address in the logical disk specified by the host (logical address) is stored. and the physical address corresponding to [when there is an updating light demand to the data stored in the disk in the past] the logical address in an address translation table 9 -- changing -- abbreviation -- it stores in the address different from the old data with other new light demand data emitted at the same stage In addition, Above LFS and the detailed explanation about WAFL are indicated by the following reference works.
[0029] : ** Reference-works 1:Ousterhout, Computer Science Division (EECS) J. -- et al. and "Beating the I/O Bottleneck:A Case for Log Structured File Systems" -- University of

California, Berkeley and UCB/CSD 88/467, October 1988.** : Reference-works 2:Seltzer, M., et al., and "An Implementation for a Log-Structured File System for UNIX", 1993 Winter USENIX, Jan.1993. [0030]
[Problem(s) to be Solved by the Invention] The following technical problems occurred in the above conventional things.
[0031] (1) : in RAID4 and RAID5 of an array disk unit, the technical problem that the overhead at the time of writing will become large occurs by existence of a light penalty.
[0032] (2) : at LFS or WAFL, the seek time to each light demand can be shortened by carrying out a buffer and carrying out the light of the light demand generated by the rose rose from a host to a disk collectively. However, you have to hold data until it stores light data in a disk after that, since a response is returned to a host when it collected and writes and data are stored in a buffer. Therefore, it must collect and write and cost must use non-volatile memory, such as comparatively high-priced NVRAM (Non Volatile Ram), for a buffer.
[0033] Moreover, if a part of file (data in a disk) is updated, since updating data are stored in a physical address different from the old data, the data of the same file are distributed and stored in a disk, and a lead performance may become bad.
[0034] this invention solves such a conventional technical problem, and aims at improving an overhead, a lead performance, etc. at the time of writing in the disk array of RAID4 and RAID5.
[0035] Moreover, when the file stored in the disk is updated, this invention stores the old data efficiently and aims at enabling it to perform efficiently the generation control of the file which holds and manages the old data for the restoration at the time of destroying a file by the failure of backup or a file.
[0036]
[Means for Solving the Problem] Drawing 1 is principle explanatory drawing of this invention. this invention was constituted as follows in order to attain the aforementioned purpose.
[0037] It has the disk array control unit 2 which performs writing / read-out control of data between two or more disk units 6 and each disk unit 6. : (1) With this disk array control unit 2 In the disk array equipment (RAID4 or RAID5) which two or more disk units 6 are made to distribute the division data of a block unit, stores, and stores in one of the disk units 6 the parity data for which it asked from these division data The parity processing section 19 which computes the exclusive OR of the data of the light demand from a host, and asks for parity data, The 1st and the 2nd conclusion writing buffer 21 and 22 which hold the data of the light demand from a host temporarily, and take double buffer composition, The parity buffer 23 which holds parity data until all the data that constitute a stripe are written in in the aforementioned conclusion writing buffer, The address translation table holding the correspondence information on the logical address specified by the host and the physical address in the disk with which data are actually stored (table in the address translation section 20), It has the logical address sorting section 18 which uses the aforementioned logical address as a key and sorts the data in the aforementioned conclusion writing buffer.
[0038] (2) : above (1) The 1st control means which are not concerned with new and updating but store the light data from a host in the conclusion writing buffer of one of the above in order of arrival in disk array equipment, When a lead demand occurs to the same disk as the disk which stores the data in the aforementioned conclusion writing buffer after the time when this conclusion writing buffer became full, If it judges whether the free area which stores all the data in the aforementioned conclusion writing buffer is near the position on the disk with which the corresponding lead data are stored and there is the aforementioned free area, he has no seeking in the free area. It has the 2nd control means which became the aforementioned full and which collect, write and store the light data of a buffer.
[0039] It has two or more disk units 6 and the disk array control unit 2 which performs writing / read-out control of data in operating each disk unit 6 in parallel. : (3) With this disk array control unit 2 In the disk array equipment (RAID4 or RAID5) which two or more disk units 6 are made to distribute the division data of a block unit, stores, and stores in one of the disk units 6 the parity data for which it asked from these division data The parity processing section 19 which was led when an updating light demand occurred from a host and which computes the parity after

updating from the data before updating and parity data in a corresponding disk, The data buffer before updating which stores the data before updating led from the disk with the address value in the disk, The data after updating in the same position on a disk, when disk media take at least 1 round, after leading the data before updating and parity from a disk, and the parity after updating are stored. When the data buffer before updating becomes full, it has the generation-control section which stores the data in a buffer in a disk.
[0040] (4) : above (3) In disk array equipment the aforementioned generation-control section The data [disk] before updating when an updating light demand occurs from a host, Lead parity and the parity after updating in the aforementioned parity processing section is made to compute. The data and parity after updating in the same position on a disk, when disk media take at least 1 round, after leading the data before updating and parity from a disk are stored. The data before updating led from the disk were stored in the aforementioned data buffer before updating with the address value in the disk, and when this data buffer before updating becomes full, it has the control means which store the data in this buffer in the continuation field in a disk.
[0041] (Operation) The operation of this invention based on the aforementioned composition is explained based on drawing 1 .
[0042] (a) : above (1) Then The 1st conclusion writing buffer 21 and the 2nd conclusion writing buffer 22, The parity processing section 19, the parity buffer 23, and an address translation table (table in the address translation section 20), If it has the logical address sorting section 18 and a light demand is emitted from a host A parity price system is carried out by the parity processing section 19, light data are gathered, it accumulates to either the 1st conclusion writing buffer 21 or the 2nd conclusion writing buffer 22, and parity data are stored in the parity buffer 23.
[0043] At this time, the address translation section 20 sets the logical address specified by the host, and the physical address corresponding to the address to an address translation table. And after sorting the data of the conclusion writing buffer which became this full by the logical address sorting section 18 at this time although the data of this conclusion writing buffer are stored in a disk if it is not the account of before but ** one conclusion writing buffer becomes full, it stores.
[0044] If it does in this way, in the disk array of RAID4 and RAID5, it will become possible to improve an overhead, a lead performance, etc. at the time of writing.
[0045] (b) : above (2) The 1st control means are not then concerned with one of conclusion writing buffers in the light data from a host at new and updating, but it stores in order of arrival. The 2nd control means and after the time when the aforementioned conclusion writing buffer became full When a lead demand occurs to the same disk as the disk which stores the data in a conclusion writing buffer, If it judges whether the free area which stores all the data in the aforementioned conclusion writing buffer is near the position on the disk with which the corresponding lead data are stored and there is the aforementioned free area The light data of the conclusion writing buffer as for which seeking became the free area nothing at the aforementioned full are stored.
[0046] If it does in this way, in the disk array of RAID4 and RAID5, it will become possible to improve an overhead, a lead performance, etc. at the time of writing.
[0047] (c) : above (3) The parity after updating is computed from the data before updating and parity data in a corresponding disk which were led when an updating light demand generated the parity processing section 19 from a host then. And the generation-control section stores in the data buffer before updating the data before updating led from the disk with the address value in the disk.
[0048] Moreover, the generation-control section stores the data in a buffer in a disk, when the data after updating in the same position on a disk when disk media take at least 1 round, and the parity after updating are stored and the data buffer before updating becomes full, after leading the data before updating, and parity from a disk.
[0049] When doing in this way and the file stored in the disk is updated, the old data are stored efficiently and the generation control of the file which holds and manages the old data for the restoration at the time of destroying a file by the failure of backup or a file can be performed efficiently.

[0050] (d) : above (4) Then the control means of the generation-control section When an updating light demand occurs from a host, the data before updating and parity are led from a disk. After making the parity after updating in the aforementioned parity processing section compute and leading the data before updating, and parity from a disk, When disk media take at least 1 round, the data and parity after updating in the same position on a disk are stored. The data before updating led from the disk are stored in the aforementioned data buffer before updating with the address value in the disk, and when this data buffer before updating becomes full, the data in this buffer are stored in the continuation field in a disk.
[0051] Like the conventional conclusion calligraphy formula (WAFL method), even when doing in this way and the part in a file is updated, since the updated data are stored in another field in a disk, they cannot be said to generate new seeking and can always store the newest data in the continuation field of a disk.
[0052] Moreover, when the file stored in the disk is updated, the old data are stored efficiently and the generation control of the file which holds and manages the old data for the restoration at the time of destroying a file by the failure of backup or a file can be performed efficiently.
[0053] (e) clear from having stated more than : -- as -- the above (1) (2) **** -- since it will end if only the conclusion writing buffer of size equal to the size of the storing data to each disk which constitutes a stripe is prepared, the amount of expensive non-volatile memory used like NVRAM which collects and writes and is prepared for buffers can be reduced Moreover, when the field which can store data without seeking in the case of storing on a disk when a lead demand takes place to the disk which it is going to store exists, the seek time in the case of a light can be shortened by storing data in the free area.
[0054] Furthermore, the data stored in the continuation field in the host are storable in a continuation field also in the conclusion writing field on a disk by collecting and writing and sorting by the demand address (logical address) from a host within a buffer (rearrangement). Therefore, the processing time at the time of leading the corresponding data can be shortened.
[0055] Moreover, the above (3) (4) Even when a part of file is updated in addition to the aforementioned point then, the newest data can always be stored in the continuation field of a disk, and a lead performance improves. moreover -- although it had managed where was updated whenever the data and the file when creating a file newly were updated in the conventional file generation control -- such [in the invention in this application] a conventional generation control -- differing -- difference with the present file -- since only data are stored in a disk, it becomes possible to restore the data before updating Therefore, control which performs restoration before what generation can be performed easily.
[0056]
[Embodiments of the Invention] Hereafter, the gestalt of implementation of invention is explained in detail based on a drawing.
[0057] **1: The explanatory drawing 2 of a system configuration and disk array equipment and 3 reference drawing 2 are explanatory drawings of a system configuration and disk array equipment, A view is a system configuration view and B view is a block diagram of disk array equipment. Drawing 2 of drawing 3 is detail drawing a part, and A view packs explanatory drawing of the address translation section, and B view, is drawn, and is explanatory drawing of a buffer.
[0058] the system shown in A view of drawing 2 -- the example of a client/server system -- it is -- a LAN top -- much client equipments 16-1 and 16-2 ... and one set of a server 15 are connected And disk array equipment 7 is connected to the server 15. in this case, the server 15 -- LAN -- minding -- two or more client equipments 16-1 and 16-2 -- each file currently used by ... collecting -- disk array equipment 7 -- holding -- each client equipment 16-1 and 16-2 -- the so-called Network Server which receives the read/write demand to the file from ... is constituted
[0059] for example, the protocol which is called NFS (Network File System) in the case of the workstation (WS) by which the aforementioned server 15 set UNIX to OS (system software) -- using -- each client equipment 16-1 and 16-2 -- when transmitting the read/write demand to FAIRU from ... in a network (it is LAN in this case), it divides into two or more comparatively small (8KB and about 32KB) data, and is made them Therefore, in case [many of] a server 15

gives a read/write demand to the disk of disk array equipment 7, it is carried out by 8KB or about 32KB of data unit.
[0060] The composition of the aforementioned disk array equipment 7 is shown in B view of drawing 2 . Although the aforementioned disk array equipment 7 is connected and applied to a server 15, if it sees from disk array equipment 7 in this case, the aforementioned server 15 will serve as a host (high order equipment). the above -- a disk array -- equipment -- seven -- RAID -- four -- or -- RAID -- five -- composition -- a disk array -- storage -- a means -- ****** -- holding -- equipment -- it is -- a disk array -- a control unit -- two -- plurality -- a disk unit -- six -- having -- the above -- a disk array -- a control unit -- two -- a disk array -- a controller -- four -- plurality -- a device -- an adapter -- (-- DA --) -- five -- having -- **** .
[0061] Moreover, the parity processing section 19 for the disk array controller 4 performing parity processing, The address translation section 20 which performs transform processing of the parity buffer 23 which stores parity data, and the logical address (address given by the host 1) and a physical address (address actually written in a disk), The 1st conclusion writing buffer 21 which stores the data transmitted by the host 1 temporarily, and the 2nd conclusion writing buffer 22, It has the above 1st, the 2nd conclusion writing buffer 21, and the logical address sorting section 18 that sorts by using the logical address to the data in 22 as a key (rearrangement processing).
[0062] That is, a server 15 holds the disk array of RAID4 or RAID5 composition as a storage means, the 1st conclusion writing buffer 21 and the 2nd conclusion writing buffer 22 take double buffer composition, and while one of the two's conclusion writing buffer is performing the data light to a disk, the light data which reach to a server 15 are stored in the conclusion writing buffer of another side.
[0063] For example, while writing the data of the 1st conclusion writing buffer 21 in the disk, when the light demand from one of client equipments reaches a server 15, the data of the light demand are stored in the 2nd conclusion writing buffer 22. Moreover, while writing the data of the 2nd conclusion writing buffer 22 in the disk, when the light demand from one of client equipments reaches a server 15, the data of the light demand are stored in the 1st conclusion writing buffer 21.
[0064] Parity data are accumulated in each stripe unit of RAID by processing by the parity processing section 19 at the aforementioned parity buffer 23. Moreover, the address translation section 20 is equipped with the address translation table as shown in A view of drawing 3 , and the information (information to which the logical address and the physical address were made to correspond) changed into the address (physical address) in the disk which collects, writes and comes out and actually stores the address (logical address) which stores in a disk the light data which the system software (OS) of a server 15 set up is written in this address translation table.

[0065] moreover, the light data d1, d2, and d3 transmitted to the 1st and 2nd conclusion writing buffer 21 and 22 from the server 15 as shown in B view of drawing 3 ... is stored And the parity processing section 19 computes the parity about these light data. For example, it can ask for the parity of the light data d1, d2, d3, and d4 by computing the exclusive OR of the data for every stripes of these.
[0066] Moreover, the parity data for which are the above, and it made and asked are stored in the parity buffer 23 by the parity processing section 19. In this case, the aforementioned parity data are held in the parity buffer 23 until all stripe data collect and write and are written in in a buffer. Furthermore, although the writing to a disk is performed in the aforementioned logical address sorting section 18 when it collected and writes and the data in a buffer become full (full), it collects at this time and writes, and after using the logical address as a key and sorting the data in a buffer, it is made to perform the writing to a disk.
[0067] **2: Explain the processing outline of disk array equipment below explanation of a processing outline. In addition, the following explanation describes a server 15 a "host." If a light demand is emitted from a host, a parity price system is carried out by the parity processing section 19, light data are gathered, it is accumulated in the 1st conclusion writing buffer 21 or

the 2nd conclusion writing buffer 22, and parity data are stored in the parity buffer 23.
[0068] In this case, although a host specifies the storing address of the aforementioned light data, as the logical address over disk accessing, the address (logical address) specified by this host is set to the address translation table of the address translation section 20, and assigns the physical address actually stored within a disk (the logical address and a physical address are made to correspond to an address translation table, and it stores in it).
[0069] Moreover, although light data are stored in the position in the disk specified by the physical address, it collects and they are written, and the size of a buffer has data size stored in each disk which constitutes RAID. Although the data sent by the host are gathered and written and are stored in order in a buffer, in that case, they perform the parity data and EXCLUSIVE OR operation in the buffer 23 for parity which were stored in the corresponding stripe, and update parity data. However, when it collects and writes and a buffer accumulates the data of the beginning of a stripe, EXCLUSIVE OR operation is not performed but stores light data also in the buffer 23 for parity.
[0070] And when it collected and writes and a buffer becomes full, after the data in this conclusion writing buffer use the logical address as a key and are sorted by the logical address sorting section 18, they are stored in one set of a disk. Storing on a disk is performed as follows. That is, when the data transmitted by the host begin to be stored in the conclusion writing buffer of another side which constitutes a TABURU buffer, a lead demand occurs on the disk which it is going to store by the time it collects and writes and a buffer becomes full, it judges whether a free area is in the same cylinder (range which can be stored without generating seeking) and there is a free area, it stores in the portion.
[0071] However, when the conclusion writing buffer of another side is full, without the ability discovering such a free area, it collects and writes to the field which has **ed in the disk, all the data in a buffer are stored, and the conclusion writing buffer for the new light demand from a host is opened. Moreover, when data are stored in all the disks that constitute a stripe, the buffer for parity is stored in a disk.
[0072] Since it will end if only the conclusion writing buffer of size equal to the size of the storing data to each disk which was described above and which constitutes a stripe (it is one stripe at d1, d2, d3, and d4 in this example) like is prepared, the amount of the non-volatile memory (this example NVRAM) which collects and writes and is prepared for buffers can be reduced. Moreover, if the field which can store data without seeking in the case of storing on a disk when a lead demand takes place to the disk which it is going to store exists, the seek time in the case of a light can be shortened by storing data in the free area.
[0073] Furthermore, the data stored in the continuation field in the host are storable in a continuation field also in the conclusion writing field on a disk by collecting and writing and sorting by the demand address (logical address) from a host within a buffer (rearrangement). Therefore, the processing time at the time of leading the corresponding data can be shortened.
[0074] **3: The explanatory drawing 4 of detailed processing is processing explanatory drawing. Hereafter, detailed processing is explained based on drawing 4 , referring to drawing 2 and drawing 3 . The light data of NFS which arrives from a network (this example LAN) are processed as follows.
[0075] the light data sent to the server 15 from the network -- updating -- it is not concerned newly but is inputted into the parity processing section 19 in order of arrival In the parity processing section 19, the parity data calculated about the stripe of the schedule in which light data are stored are read from the parity buffer 23, EXCLUSIVE OR operation with light data is performed, parity is updated, and it stores in the original parity storage region in the parity buffer 23 here.
[0076] updating after gathering the aforementioned light data, writing and being given the physical address in a buffer -- it is not concerned newly, but it collects in order of arrival, writes, and is stored in a buffer The size of a conclusion writing buffer has a size stored in the continuation field of one disk which constitutes RAID, and the storing address (logical address) which the system software (OS) of a server 15 set up is set to an address translation table with the address (physical address) which collected and wrote and was stored in the buffer.

[0077] And when it collected and writes and a buffer becomes full (full) by light data, it collects and writes and the data storage to a buffer is started to the conclusion writing buffer of another side which constitutes a double buffer. And the light [which became full by light data] data in which the buffer was stored by the logical address sorting section 18 by collecting and writing are sorted with each logical address value. Then, it stores in a disk.
[0078] for example, the case where light data are first stored in the 1st conclusion writing buffer 21 -- this -- if the 1st conclusion writing buffer 21 becomes full by light data, the light data stored in this 1st conclusion writing buffer 21 are stored in the disk sorted with each logical address value At this time, the light data transmitted by the host are accumulated at the 2nd conclusion writing buffer 22.
[0079] The light demand to the file from each client equipment is transmitted to a server 15 as a light demand of two or more NFS to the continuous field. However, in a server 15, in order to receive the demand from two or more client equipments, the light demand which the demand from other client equipments might interrupt and continued is because it is not not necessarily access to a continuation field.
[0080] Furthermore, the writing to a disk is stored in the field, when a lead demand on the same disk occurs and the continuation field which collects and writes to near (inside of the range without seek operation, for example, the same cylinder), and can store the data in a buffer is discovered. By doing in this way, the time which seeking required in the case of the data storage to a disk takes is omissible.
[0081] Under the present circumstances, the physical address in an address translation table is changed into the address in a disk. By the time other conclusion writing buffers which constitute a double buffer will become full by light data, when such a field cannot be discovered, it seeks, and collects and writes to a continuation field, and the data in a buffer are stored. The data in the parity buffer 23 are similarly stored in a disk.
[0082] In addition, the data which collected, wrote and were stored in the buffer are stored in the field when a free area is near [in the disk with which lead data are stored on the occasion of the lead of other data]. Therefore, although the storing positions of the data which constitute one stripe may differ between disks, the correspondence table showing where a logical stripe is actually stored in each disk can be prepared, and it can be coped with by referring to the table.
[0083] the example of drawing 4 -- the light data d1, d2, and d3 of the light demand from a network -- although ... is transmitted as block data of nKB, respectively, the light data of nKB of d1 are [the data from client equipment 16-2 and the light data of nKB of d3 of the data from client equipment 16-1 and the light data of nKB of d2] data from client equipment 16-3, for example
[0084] Moreover, in this example, it is the example which made the disk for data four sets, and the data of a stripe which consist of d1, d2, d3, and d4 are stored in one set (a part for for example, two to 1 light unit =3 truck) of a disk. Moreover, Above nKB is data for 128 sectors of a disk, this data of nKB is made into a unit and the logical address is specified.
[0085] In addition, although it collected and wrote and the buffer of double composition explained the buffer in the aforementioned example, even if it uses three or more conclusion writing buffers, it can carry out. However, since the number of buffers, such as expensive NVRAM, increases in this case and it leads to the part and a cost rise, it is optimal to use two conclusion writing buffers.
[0086] (others -- explanation of an example)
**1: Explanatory-drawing 5 reference drawing 5 of the equipment of other examples is the equipment block diagram of other examples. In the disk array equipment of RAID4 or RAID5, this example is an example which performs a generation control, and is constituted like drawing 5 .
[0087] a disk array -- equipment -- seven -- RAID -- four -- or -- RAID -- five -- composition -- a disk array -- storage -- a means -- ****** -- holding -- equipment -- it is -- a disk array -- a control unit -- two -- plurality -- a disk unit -- six -- having -- the above -- a disk array -- a control unit -- two -- a disk array -- a controller -- four -- plurality -- a device -- an adapter -- (-- DA --) -- five -- having -- **** . Moreover, the disk array controller 4 is equipped with the parity processing section 19 for performing parity processing,

the parity buffer 23 which stores parity data, the generation-control section 25, and the front [updating] data baffy 26 grade.
[0088] The aforementioned generation-control section 25 performs the generation control of data, a parity operation is made to perform, or when it takes out directions to the parity processing section 19 and receives updating data from a host, it stores the data before updating in the data buffer 26 before updating. In addition, other composition is the same as the disk array equipment shown in drawing 2 .
[0089] **2: When the updating light demand from a host occurs, after the explanation generation-control section 25 of processing reads the data and parity before updating from the disk of a disk unit 6 and makes the parity after updating in the parity processing section 19 calculate, it stores the data before updating in the data buffer 26 before updating, and stores the data and parity after updating in a disk.
[0090] The data one generation before the data updated for every fixed period (the address on a disk is included) are held by setting to the data buffer 26 before updating the address made equivalent to the address on the disk with which the data and data before updating were stored. And when a host wants to refer to the data one generation before a certain file, the field where the data before updating are stored is searched based on the address value to which the data belonging to the file are set.
[0091] If the updating data from a host reach the generation-control section 25, the generation-control section 25 will read the data and parity before updating from a disk. And in the parity processing section 19, it creates from the data after the data before updating new parity, parity, and updating, and stores in the parity buffer 23.
[0092] Then, when the generation-control section 25 reads the data before updating and disk media carry out it at least 1 round, it stores the data after updating in the same position. Similarly, while disk media take at least 1 round, the parity after updating also makes the parity after updating in the parity processing section 19 calculate, and is stored in the same position. However, when the lead of the data before updating becomes slow, disk media may rotate 1 round or more.
[0093] The data before updating are stored in the data puffer 26 before updating with the address value which stored the data. The data in the front [updating] data buffer 26 are stored into a disk by the instruction from a host for every fixed period.
[0094] Even when making it above and a part of file is updated, the newest data can always be stored in the continuation field of a disk, and a lead performance improves. or [moreover, / the data when creating a file newly like the conventional file generation control, and / that where was updated whenever the file was updated] -- one by one -- not memorizing -- difference with the present file -- by storing only data, the data before updating can be restored and control which performs restoration before what generation can be performed easily
[0095]
[Effect of the Invention] As explained above, according to this invention, there are the following effects.
[0096] (1) : in a claim 1, if it has the 1st conclusion writing buffer, the 2nd conclusion writing buffer, the parity processing section, a parity buffer, an address translation table, and the logical address sorting section and a light demand is emitted from a host, a parity price system is carried out by the parity processing section, gather light data, and it is accumulated in the 1st conclusion writing buffer or the 2nd conclusion writing buffer, and store parity data in a parity buffer.
[0097] At this time, the address translation section sets the logical address specified by the host, and the physical address corresponding to the address to an address translation table. And after sorting the data of the conclusion writing buffer which became this full by the logical address sorting section at this time although the data of this conclusion writing buffer are stored in a disk if the conclusion writing buffer of one of the above becomes full, it stores.
[0098] If it does in this way, in the disk array of RAID4 and RAID5, it will become possible to improve an overhead, a lead performance, etc. at the time of writing.
[0099] (2) : in a claim 2, the 1st control means are not concerned with new and updating, but

store the light data from a host in one of conclusion writing buffers in order of arrival. The 2nd control means and after the time when the aforementioned conclusion writing buffer became full When a lead demand occurs to the same disk as the disk which stores the data in a conclusion writing buffer, If it judges whether the free area which stores all the data in the aforementioned conclusion writing buffer is near the position on the disk with which the corresponding lead data are stored and there is the aforementioned free area, he has no seeking in the free area. The light data of the conclusion writing buffer which became the aforementioned full are stored.
[0100] If it does in this way, in the disk array of RAID4 and RAID5, it will become possible to improve an overhead, a lead performance, etc. at the time of writing.
[0101] (3) : in a claim 3, the parity processing section computes the parity after updating from the data before updating and parity data in a corresponding disk which were led when an updating light demand occurred from a host. And the generation-control section stores in the data buffer before updating the data before updating led from the disk with the address value in the disk.
[0102] Moreover, the generation-control section stores the data in a buffer in a disk, when the data after updating in the same position on a disk when disk media take at least 1 round, and the parity after updating are stored and the data buffer before updating becomes full, after leading the data before updating, and parity from a disk.
[0103] When doing in this way and the file stored in the disk is updated, the old data are stored efficiently and the generation control of the file which holds and manages the old data for the restoration at the time of destroying a file by the failure of backup or a file can be performed efficiently.
[0104] In a claim 4 : (4) The control means of the generation-control section When an updating light demand occurs from a host, lead the data before updating and parity from a disk, and the parity after updating is computed. When disk media take at least 1 round after leading the data before updating, and parity from a disk The data before updating which stored the data and parity after updating in the same position on a disk, and were led from the disk with the address value in the disk When it stores in the aforementioned data buffer before updating and this data buffer before updating becomes full, the data in this buffer are stored in the continuation field in a disk.
[0105] Like the conventional conclusion calligraphy formula (WAFL method), even when doing in this way and the part in a file is updated, since the updated data are stored in another field in a disk, they cannot be said to generate new seeking and can always store the newest data in the continuation field of a disk.
[0106] Moreover, when the file stored in the disk is updated, the old data are stored efficiently and the generation control of the file which holds and manages the old data for the restoration at the time of destroying a file by the failure of backup or a file can be performed efficiently.
[0107] (5) : since it will end with claims 1 and 2 if only the conclusion writing buffer of size still more nearly equal to the size of the storing data to each disk which constitutes a stripe is prepared, the amount of expensive non-volatile memory used like NVRAM which collects and writes and is prepared for buffers can be reduced.
[0108] Moreover, when the field which can store data without seeking in the case of storing on a disk when a lead demand takes place to the disk which it is going to store exists, the seek time in the case of a light can be shortened by storing data in the free area.
[0109] Furthermore, the data stored in the continuation field in the host are storable in a continuation field also in the conclusion writing field on a disk by collecting and writing and sorting by the demand address (logical address) from a host within a buffer (rearrangement). Therefore, the processing time at the time of leading the corresponding data can be shortened.
[0110] (6) : in claims 3 and 4, further, even when a part of file is updated, the newest data can always be stored in the continuation field of a disk, and a lead performance improves. moreover -- although it had managed where was updated whenever the data and the file when creating a file newly were updated in the conventional file generation control -- such [in the invention in this application] a conventional generation control -- differing -- difference with the present file -- since only data are stored in a disk, it becomes possible to restore the data before updating

Therefore, control which performs restoration before what generation can be performed easily.
[0111] (7) : in the disk array of RAID4 and RAID5, an overhead, a lead performance, etc. at the time of writing are improvable. Moreover, when the file stored in the disk is updated, the old data are stored efficiently and the generation control of the file which holds and manages the old data for the restoration at the time of destroying a file by the failure of backup or a file can be performed efficiently.

[Translation done.]

* NOTICES *

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1] It is principle explanatory drawing of this invention.
[Drawing 2] It is explanatory drawing of a system configuration and disk array equipment in the gestalt of operation of this invention.
[Drawing 3] a part of drawing 2 -- it is detail drawing
[Drawing 4] It is processing explanatory drawing in the gestalt of operation of this invention.
[Drawing 5] It is the equipment block diagram of other examples of this invention.
[Drawing 6] It is explanatory drawing of conventional disk array equipment.
[Drawing 7] It is level explanatory drawing of RAID.
[Drawing 8] It is explanatory drawing of WAFL.
[Description of Notations]
1 Host (Host Computer)
2 Disk Array Control Unit
3 Host Adaptor
4 Disk Array Controller
5-1 - 5 and 5-n Device adapter
6-1 - 6 and 6-n Disk unit
7 Disk Array Equipment
8 Data Buffer
9 Address Translation Table
10 Conclusion Writing Buffer
15 Server
16 Client Equipment
18 Logical Address Sorting Section
19 Parity Processing Section
20 Address Translation Section
21 1st Conclusion Writing Buffer
22 2nd Conclusion Writing Buffer

[Translation done.]

(19)日本国特許庁（ＪＰ）

(12) **公開特許公報**（Ａ）

(11)特許出願公開番号
特開2001－51806
（P2001－51806A）

(43)公開日　平成13年２月23日(2001.2.23)

(51)Int.Cl.[7]	識別記号	ＦＩ	テーマコード（参考）
Ｇ０６Ｆ　3/06	３０５	Ｇ０６Ｆ　3/06	３０５Ｃ　５Ｂ０６５
	５４０	-	５４０

審査請求　未請求　請求項の数４　ＯＬ　（全 14 頁）

(21)出願番号　特願平11－222078

(22)出願日　平成11年８月５日(1999.8.5)

(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号

(72)発明者　太田　善之
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

(72)発明者　西川　克彦
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

(74)代理人　100096530
弁理士　今村　辰夫　（外２名）

最終頁に続く

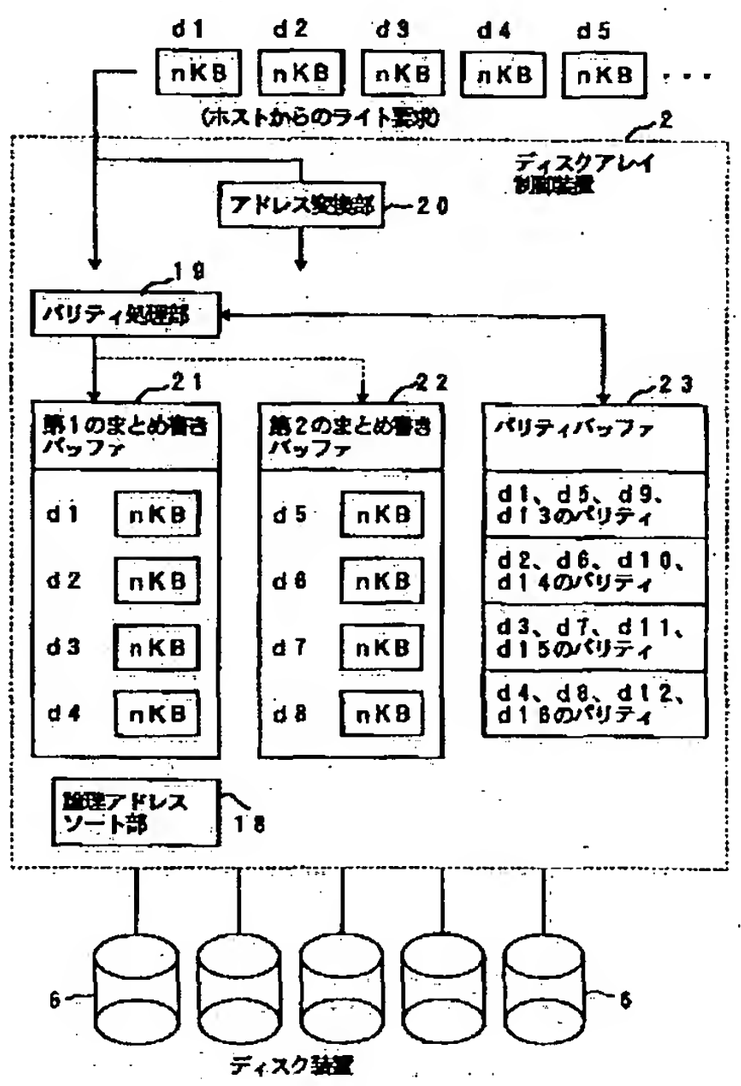
(54)【発明の名称】　ディスクアレイ装置

(57)【要約】

【課題】本発明はディスクアレイ装置に関し、ＲＡＩＤ４、ＲＡＩＤ５のディスクアレイにおいて、書き込み時のオーバーヘッドやリード性能などを改善する。また、ファイルの世代管理を効率良く行えるようにする。

【解決手段】ホストからのライト要求のデータ同士の排他的論理和を算出してパリティデータを求めるパリティ処理部１９と、ホストからのライト要求のデータを一時的に保持し、ダブルバッファ構成をとる第１、第２のまとめ書きバッファ２１、２２と、まとめ書きバッファ内にストライプを構成する全てのデータが書き込まれるまでパリティデータを保持するパリティバッファと、論理アドレスと物理アドレスとの対応を保持するアドレス変換テーブルと、まとめ書きバッファ内のデータを、論理アドレスをキーにしてソートする論理アドレスソート部１８とを備えた。

本発明の原理説明図



【特許請求の範囲】
【請求項１】複数のディスク装置と、各ディスク装置との間でデータの書き込み／読み出し制御を行うディスクアレイ制御装置を備え、該ディスクアレイ制御装置により、ブロック単位の分割データを複数のディスク装置に分散させて格納し、該分割データから求めたパリティデータをいずれかのディスク装置に格納するディスクアレイ装置において、
ホストからのライト要求のデータ同士の排他的論理和を算出してパリティデータを求めるパリティ処理部と、
ホストからのライト要求のデータを一時的に保持し、ダブルバッファ構成をとる第１、第２のまとめ書きバッファと、
前記まとめ書きバッファ内に、ストライプを構成する全てのデータが書き込まれるまでパリティデータを保持するパリティバッファと、
ホストから指定された論理アドレスと、実際にデータが格納されるディスク内の物理アドレスとの対応情報を保持するアドレス変換テーブルと、
前記まとめ書きバッファ内のデータを、前記論理アドレスをキーにしてソートする論理アドレスソート部を備えている、
ことを特徴とするディスクアレイ装置。
【請求項２】ホストからのライトデータを、前記いずれか一方のまとめ書きバッファに、新規、更新に関わらず到着順に格納する第１の制御手段と、
該まとめ書きバッファがフルになった時刻以降に、前記まとめ書きバッファ内のデータを格納するディスクと同一ディスクに対してリード要求が発生した際、該当するリードデータが格納されているディスク上の位置の近傍に、前記まとめ書きバッファ内の全データを格納する空き領域があるかどうかを判断し、前記空き領域があれば、その空き領域にシーク無しで、前記フルになったまとめ書きバッファのライトデータを格納する第２の制御手段を備えている、
ことを特徴とする請求項１記載のディスクアレイ装置。
【請求項３】複数のディスク装置と、各ディスク装置を並列的に動作させることでデータの書き込み／読み出し制御を行うディスクアレイ制御装置を備え、該ディスクアレイ制御装置により、ブロック単位の分割データを複数のディスク装置に分散させて格納し、該分割データから求めたパリティデータをいずれかのディスク装置に格納するディスクアレイ装置において、
ホストから更新ライト要求が発生した際にリードされた、対応するディスク内の更新前データとパリティデータから、更新後のパリティを算出するパリティ処理部と、
ディスクからリードされた更新前データを、そのディスク内のアドレス値と共に格納する更新前データバッファと、
ディスクから更新前データとパリティをリードした後、ディスク媒体が少なくとも１周した際にディスク上の同一位置に更新後のデータ及び更新後のパリティを格納し、更新前データバッファがフルになった際に該バッファ内のデータをディスクに格納する世代管理部を備えている、
ことを特徴とするディスクアレイ装置。
【請求項４】前記世代管理部は、
ホストから更新ライト要求が発生した際に、ディスクから対応する更新前データとパリティをリードして前記パリティ処理部に更新後のパリティを算出させ、ディスクから更新前データとパリティをリードした後、ディスク媒体が少なくとも１周した際に、ディスク上の同一位置に更新後のデータ及びパリティを格納し、ディスクからリードした更新前データを、そのディスク内のアドレス値と共に前記更新前データバッファに格納し、該更新前データバッファがフルになった際に該バッファ内のデータをディスク内の連続領域に格納する制御手段を備えている、
ことを特徴とする請求項３記載のディスクアレイ装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、複数のディスク装置と前記ディスク装置を並列的に動作させてデータの読み出し／書き込み制御を行うディスクアレイ制御装置を備えたＲＡＩＤレベル４又はＲＡＩＤレベル５のディスクアレイ装置に関する。
【０００２】
【従来の技術】以下、従来例を説明する。
【０００３】§１：ディスクアレイ装置の説明・・・図６参照
図６は従来のディスクアレイ装置の説明図である。ディスクアレイ装置は内蔵した複数の磁気ディスク装置（ハードディスク装置）を並列動作させることで、データの読み出し／書き込み速度の高速化を図り、かつ冗長構成の導入によって信頼性を向上させた外部記憶装置、或いは補助記憶装置である。なお、以下の説明では、前記磁気ディスク装置（又はハードディスク装置）を単に「ディスク装置」と記す。
【０００４】図６に示したように、ディスクアレイ装置は、ディスクアレイ制御装置２と、複数のＲＡＩＤ（詳細は後述する）を構成するディスク装置６－１、６－２、６－３・・・６－m、６－ｎ（ｎ＝m＋１）で構成されている。また、ディスクアレイ制御装置２には、ホストアダプタ３と、ディスクアレイコントローラ４と、複数のデバイスアダプタ（ＤＡ）５－１、５－２、５－３・・・５－m、５－ｎが設けてある。そして、デバイスアダプタ５－１～５－ｎには、それぞれ各ディスク装置６－１～６－ｎが接続されている。
【０００５】前記ディスクアレイ装置はホスト１に接続

されて運用されるが、この場合、ホスト１とディスクアレイ制御装置２のホストアダプタ３間をインターフェースケーブル（例えば、ＳＣＳＩ－２用ケーブル）により接続する。ホストアダプタ３はホスト１に対するインターフェース制御を行うものであり、ディスクアレイコントローラ４は、データのリード／ライト時の各種制御等を行うものである。デバイスアダプタ（ＤＡ）５－１～５－ｎは、データのリード／ライト時にディスクアレイコントローラ４の指示によりディスク装置６－１～６－ｎに対する制御を行うものである。

【０００６】ディスクアレイ装置はホスト１から見ると１台のディスク装置に見える。このディスクアレイ装置では、例えば、ホストアダプタ３がホスト１から送られたデータを受信すると、そのデータをディスクアレイコントローラ４へ送る。そして、ディスクアレイコントローラ４は、例えば、前記データを複数のデータに分割し、デバイスアダプタ５－１～５－ｍを介して複数のディスク装置６－１～６－ｍ（データ用ディスク装置）に格納すると共に、前記データに対するパリティデータを作成し、デバイスアダプタ５－ｎを介して残りの１台のディスク装置６－ｎ（パリティ用ディスク装置）に格納する。

【０００７】このようにディスクアレイ装置は、大きなサイズのデータを複数のディスク装置に同時に書き込んだり、或いは複数のディスク装置から同時に読み出したりすることで１台のディスク装置よりもリード／ライトの高速化を実現し、かつデータの信頼性を向上させることができる。従って、装置の高性能化を図ることができるものである。

【０００８】§２：ＲＡＩＤの説明

前記ディスクアレイ装置は、複数のディスク装置（ハードディスク装置）を使用することにより、単独のディスク装置よりも高い信頼性と性能を実現する方式である。これは、１９８７年に米国のカリフォルニア大学バークレイ校のデビッド．Ａ．パターソン(David.A.Patterson) 教授らが提唱したＲＡＩＤ（Redundant Arrays of Inexpensive Disks ）と呼ばれるものである。すなわち、前記ＲＡＩＤは前記デビッド．Ａ．パターソン教授らの論文に由来する呼び方である。

【０００９】前記のように、高速に大量のデータを多くのディスク装置にアクセスし、ディスク故障時におけるデータの冗長性を実現するディスクアレイ方式ＲＡＩＤは、１から５までのレベル（以下、前記レベルをＲＡＩＤ１、ＲＡＩＤ２、ＲＡＩＤ３、ＲＡＩＤ４、ＲＡＩＤ５とも記す）に分類されており、レベル３からレベル５（ＲＡＩＤ３、ＲＡＩＤ４、ＲＡＩＤ５）には、ディスク装置の故障時にデータを回復させるためのパリティデータを保持する。

【００１０】前記ＲＡＩＤのレベルの中でも、レベル４（ＲＡＩＤ４）、レベル５（ＲＡＩＤ５）では、複数の同時読み出しが可能であり、更に、レベル５（ＲＡＩＤ５）においては、パリティを格納するディスク装置を固定しないことで、複数の同時書き込みをも可能としており、大量のトランザクション処理において効果を発揮する。

【００１１】§３：ＲＡＩＤのレベルの説明・・・図７参照

図７はＲＡＩＤのレベル説明図であり、①図はＲＡＩＤ３、４のデータ転送（書き込み）、②図はＲＡＩＤ５のデータ転送（書き込み）を示した図である。なお、図７において、３はホストアダプタ、４はディスクアレイコントローラ、Ａ、Ｂ、Ｃ、Ｄ、Ｅはそれぞれデバイスアダプタ（ＤＡ）、６－１～６－５はディスク装置である。

【００１２】(1) ：ＲＡＩＤ１の説明

ＲＡＩＤ１は、ディスク装置を二重化した、いわゆるミラードディスク構成である。すなわち、２つのディスク装置に同じデータが書かれる。ディスク装置のコストが２倍かかるが、最もシンプルで実績もある。性能に関しては、データの書き込み時は、２つのディスク装置への書き込み完了を待つために実行時間が少し延びるが、データの読み取り時は、２つのディスク装置のどちらかが空いていれば原理上は実行可能なので、性能向上となる。

【００１３】(2) ：ＲＡＩＤ２の説明

ＲＡＩＤ２は、入力データを分割し（ストライピング）、インターリーブをかけて複数のディスク装置に分割して格納するものである。この場合、前記分割したデータに対するエラー訂正符号を格納するディスク装置を冗長ディスク装置とし、前記冗長ディスク装置のデータをハミングコードとしたものである。

【００１４】ＲＡＩＤ２では、例えば、データ格納用のディスク装置が４台の場合、エラー訂正符号用のディスク装置は３台必要である。ディスク装置の特徴としては、複数のディスク装置障害でもデータが失われないことである。そのために、前記のように、複数の冗長ディスク装置を必要とし、正味のデータ量が少なくなるという欠点がある。

【００１５】(3) ：ＲＡＩＤ３の説明・・・図７の①図参照

ＲＡＩＤ３は、入力データを分割し（ストライピング）、インターリーブをかけて分割したデータを複数のディスク装置に分散して格納するものである。この場合、前記分割したデータに対するエラー訂正符号を格納するディスク装置を冗長ディスク装置とし、前記冗長ディスク装置のデータをパリティデータとしたものである。

【００１６】このため、ＲＡＩＤ３ではデータ用のディスク装置の数に関わらず、前記冗長ディスク装置（パリティデータ用ディスク装置）の台数は１台で済む。例え

ば、ＲＡＩＤ３では、図７の①図に示したように、データを複数のディスク装置６－１～６－４に対して並列的にリード／ライトするもので、１台のパリティ用ディスク装置６－５を持つことを特徴としている。

【００１７】データ転送速度は、並列度によりＮ倍（Ｎ：並列動作させるデータ用ディスク装置の台数）となる。また、１台のディスク装置に障害が発生しても、性能低下はなく、原理的にはデータ転送中に障害が発生しても、そのまま動作することも可能である。

【００１８】(4)：ＲＡＩＤ４の説明・・・図７の①図参照

ＲＡＩＤ４は、ＲＡＩＤ３におけるストライピングユニット（分割単位）をセクタ単位（１又は数セクタ）としたものである。すなわち、前記ＲＡＩＤ３では、複数のディスク装置を束ねてアクセスするため、ある程度大きな単位でアクセスする時は性能上の効果が見られるが、小さなデータ量のアクセスでは、個々のディスク装置にアクセスを分散させた方が性能上有利である。

【００１９】このことから、ＲＡＩＤ４では、データは各ディスク装置に独立に書くが、パリティディスク装置を持ち、各ディスク装置の対応するビットから生成されたパリティを格納する。そのため、１台のディスク装置の障害でも、データが失われない。欠点は、データ更新時に、必ずデータディスク装置とパリティディスク装置の両方をアクセスしなければならないことで、この場合のパリティディスク装置がボトルネックに成りやすい。

【００２０】(5)：ＲＡＩＤ５の説明・・・図７の②図参照

ＲＡＩＤ５は、ＲＡＩＤ４の欠点であったパリティディスク装置へのアクセス集中を解消するために、パリティデータを各ディスク装置６－１～６－５に分散させたものである。しかし、依然としてデータの更新時には２つのディスクアレイ装置をアクセスしなければならないし、パリティデータは、原理上、旧データとの差分をとらなければ生成できないので、リード（読み出し）、データ生成、ライト（書き込み）というプロセスを取る必要がある。

【００２１】更に、パリティデータによって、データ喪失は事実上なくなったと言えるが、１台のディスクアレイ装置が故障した場合は、少なくとも元の装置性能を維持することができない。故障したディスクアレイ装置のデータを作り出すのに、残ったディスクアレイ装置のデータを総動員するからである。従って、ノンストップシステムのように、ディスクアレイ装置の障害でシステムを止めたくないものには向かない。

【００２２】前記ＲＡＩＤ５のデータ転送（書き込み）の１例は図７の②図に示した通りである。②図において、Ｐ０、Ｐ１、Ｐ２、Ｐ３はパリティであり、パリティ生成のセクタグループ毎にパリティを格納するディスク装置を異ならせている。なお、ＲＡＩＤ５でもＲＡＩＤ４と同様にストライピングユニットはセクタ単位（１又は数セクタ）である。

【００２３】(6)：その他の説明

前記ＲＡＩＤ４及びＲＡＩＤ５として規定されているディスクアレイ装置のアクセス方法の中で、例えば、データをブロック単位で分割し、その分割したデータをアレイを構成する各ディスクに分散させて記憶するものがある。この場合、１つのディスクの故障の際に、データを復元するため、分散させて記憶したデータの排他的論理和を計算し、パリティデータとして別ディスクに記憶している。

【００２４】このＲＡＩＤ４、５のディスクアレイに対し、データ更新のための書き込みを行う場合、該当する旧データとそれに対応するパリティデータを一度読み出し、更新データを書き込むと共に、新たなパリティデータを計算して書き込みを行う必要がある。この為、通常の単体ディスクに対するデータ書き込みに比べ、余分なディスクアクセスが必要になる（ライトペナルティ）。

【００２５】§４：ＬＦＳ、及びＷＡＦＬの説明・・・図８参照

図８はＷＡＦＬの説明図である。前記のように、ＲＡＩＤアレイでは、並列アクセスにより外部記憶装置のスループットの向上を図っているが、ＲＡＩＤ４、ＲＡＩＤ５では、上記のライトペナルティ（前記の余分なディスクアクセス）の存在により書き込み時のオーバーヘッドが大きくなってしまうという問題がある。これに対して、ＬＦＳ（Log-Structured File System）、或いはＷＡＦＬ（Write Anywhere File Layout）と呼ばれる技術がある。

【００２６】前記ＬＦＳ及びＷＡＦＬでは、各々のライト要求のデータ、例えば、ｎＫＢ単位のライトデータｄ１、ｄ２、ｄ３、ｄ４、ｄ５、ｄ６・・・を、新規、更新に関わらず、一旦ディスクコントローラ内のまとめ書きバッファ１０に格納し、ホストへレスポンスを返す。

【００２７】その後、まとめ書きバッファ１０がフル（満杯）になった時点で、一括してディスク内の連続領域に格納する。この場合、パリティデータは、まとめ書きバッファ１０に次々と格納されていくｎＫＢ単位のライトデータｄ１、ｄ２、ｄ３、ｄ４、ｄ５、ｄ６・・・間で作成する。

【００２８】また、アドレス変換テーブル９内には、ホストから指定された論理的なディスク内のアドレス（論理アドレス）と実際にデータが記憶される物理アドレスとの対応を格納しておく。そして、過去にディスクに格納したデータに対する更新ライト要求があった場合には、アドレス変換テーブル９内の論理アドレスに対応する物理アドレスを変更し、略同じ時期に発せられる他の新たなライト要求データと共に、旧データとは別のアドレスに格納する。なお、前記ＬＦＳ、及びＷＡＦＬに関する詳細な説明は、次の参考資料に記載されている。

【００２９】①：参考資料１：Ousterhout, J.,et al.,"Beating the I/O Bottleneck:A Case for Log Structured File Systems," Computer Science Division(EECS),University of California, Berkeley, UCB/CSD 88/467, October 1988.
②：参考資料２：Seltzer,M.,et al.,"An Implementation for a Log-Structured File System for UNIX," 1993 Winter USENIX,Jan.1993.
【００３０】
【発明が解決しようとする課題】前記のような従来のものにおいては、次のような課題があった。
【００３１】(1)：アレイディスク装置のＲＡＩＤ４、ＲＡＩＤ５では、ライトペナルティの存在により、書き込み時のオーバーヘッドが大きくなってしまう、という課題がある。
【００３２】(2)：ＬＦＳ、或いはＷＡＦＬでは、ホストからバラバラに発生されるライト要求をバッファし、一括してディスクへライトすることで、各々のライト要求に対するシーク時間を短縮することができる。しかし、まとめ書きバッファにデータを格納した時点でホストへレスポンスを返してしまうので、その後、ライトデータをディスクに格納するまで、データを保持しなければならない。従って、まとめ書きバッファにはコストが割高なＮＶＲＡＭ（Non Volatile Ram）等の不揮発性メモリを使用しなければならない。
【００３３】また、ファイル（ディスク内データ）の一部を更新すると、更新データは旧データとは別の物理アドレスに格納されるため、同一ファイルのデータがディスク内に分散されて格納され、リード性能が悪くなる可能性がある。
【００３４】本発明は、このような従来の課題を解決し、ＲＡＩＤ４、ＲＡＩＤ５のディスクアレイにおいて、書き込み時のオーバーヘッドやリード性能などを改善することを目的とする。
【００３５】また、本発明は、ディスクに格納されているファイルが更新された際に、旧データを効率良く格納し、バックアップやファイルの操作ミスでファイルを破壊した際の復旧のために、旧データを保持し管理するファイルの世代管理を効率良く行えるようにすることを目的とする。
【００３６】
【課題を解決するための手段】図１は本発明の原理説明図である。本発明は前記の目的を達成するため、次のように構成した。
【００３７】(1)：複数のディスク装置６と、各ディスク装置６との間でデータの書き込み／読み出し制御を行うディスクアレイ制御装置２を備え、該ディスクアレイ制御装置２により、ブロック単位の分割データを複数のディスク装置６に分散させて格納し、該分割データから求めたパリティデータをいずれかのディスク装置６に格納するディスクアレイ装置（ＲＡＩＤ４、又はＲＡＩＤ５）において、ホストからのライト要求のデータ同士の排他的論理和を算出してパリティデータを求めるパリティ処理部１９と、ホストからのライト要求のデータを一時的に保持し、ダブルバッファ構成をとる第１、第２のまとめ書きバッファ２１、２２と、前記まとめ書きバッファ内に、ストライプを構成する全てのデータが書き込まれるまでパリティデータを保持するパリティバッファ２３と、ホストから指定された論理アドレスと、実際にデータが格納されるディスク内の物理アドレスとの対応情報を保持するアドレス変換テーブル（アドレス変換部２０内のテーブル）と、前記まとめ書きバッファ内のデータを、前記論理アドレスをキーにしてソートする論理アドレスソート部１８とを備えている。
【００３８】(2)：前記(1)のディスクアレイ装置において、ホストからのライトデータを、前記いずれか一方のまとめ書きバッファに、新規、更新に関わらず到着順に格納する第１の制御手段と、該まとめ書きバッファがフルになった時刻以降に、前記まとめ書きバッファ内のデータを格納するディスクと同一ディスクに対してリード要求が発生した際、該当するリードデータが格納されているディスク上の位置の近傍に、前記まとめ書きバッファ内の全データを格納する空き領域があるかどうかを判断し、前記空き領域があれば、その空き領域にシーク無しで、前記フルになったまとめ書きバッファのライトデータを格納する第２の制御手段を備えている。
【００３９】(3)：複数のディスク装置６と、各ディスク装置６を並列的に動作させることでデータの書き込み／読み出し制御を行うディスクアレイ制御装置２を備え、該ディスクアレイ制御装置２により、ブロック単位の分割データを複数のディスク装置６に分散させて格納し、該分割データから求めたパリティデータを、いずれかのディスク装置６に格納するディスクアレイ装置（ＲＡＩＤ４、又はＲＡＩＤ５）において、ホストから更新ライト要求が発生した際にリードされた、対応するディスク内の更新前データとパリティデータから、更新後のパリティを算出するパリティ処理部１９と、ディスクからリードされた更新前データを、そのディスク内のアドレス値と共に格納する更新前データバッファと、ディスクから更新前データとパリティをリードした後、ディスク媒体が少なくとも１周した際にディスク上の同一位置に更新後のデータ及び更新後のパリティを格納し、更新前データバッファがフルになった際にバッファ内のデータをディスクに格納する世代管理部を備えている。
【００４０】(4)：前記(3)のディスクアレイ装置において、前記世代管理部は、ホストから更新ライト要求が発生した際に、ディスクから対応する更新前データと、パリティをリードして前記パリティ処理部に更新後のパリティを算出させ、ディスクから更新前データとパリティをリードした後、ディスク媒体が少なくとも１周した

際にディスク上の同一位置に更新後のデータ及びパリティを格納し、ディスクからリードした更新前データを、そのディスク内のアドレス値と共に前記更新前データバッファに格納し、該更新前データバッファがフルになった際に該バッファ内のデータをディスク内の連続領域に格納する制御手段を備えている。

【００４１】（作用）前記構成に基づく本発明の作用を、図１に基づいて説明する。

【００４２】(a)：前記(1)では、第１のまとめ書きバッファ２１と、第２のまとめ書きバッファ２２と、パリティ処理部１９と、パリティバッファ２３と、アドレス変換テーブル（アドレス変換部２０内のテーブル）と、論理アドレスソート部１８とを備え、ホストからライト要求が発せられると、パリティ処理部１９によりパリティ計算し、ライトデータはまとめて、第１のまとめ書きバッファ２１、又は第２のまとめ書きバッファ２２のいずれか一方に蓄積し、パリティデータはパリティバッファ２３に格納する。

【００４３】この時、アドレス変換部２０は、ホストが指定した論理アドレスと、そのアドレスに対応した物理アドレスとをアドレス変換テーブルにセットする。そして、前記いずか一方のまとめ書きバッファがフルになったら、該まとめ書きバッファのデータをディスクへ格納するが、この時、該フルになったまとめ書きバッファのデータを、論理アドレスソート部１８によりソートしてから格納する。

【００４４】このようにすれば、ＲＡＩＤ４、ＲＡＩＤ５のディスクアレイにおいて、書き込み時のオーバーヘッドやリード性能などを改善することが可能になる。

【００４５】(b)：前記(2)では、第１の制御手段は、ホストからのライトデータを、いずれか一方のまとめ書きバッファに、新規、更新に関わらず到着順に格納する。そして、第２の制御手段は、前記まとめ書きバッファがフルになった時刻以降に、まとめ書きバッファ内のデータを格納するディスクと同一ディスクに対してリード要求が発生した際、該当するリードデータが格納されているディスク上の位置の近傍に、前記まとめ書きバッファ内の全データを格納する空き領域があるかどうかを判断し、前記空き領域があれば、その空き領域にシーク無しで前記フルになったまとめ書きバッファのライトデータを格納する。

【００４６】このようにすれば、ＲＡＩＤ４、ＲＡＩＤ５のディスクアレイにおいて、書き込み時のオーバーヘッドやリード性能などを改善することが可能になる。

【００４７】(c)：前記(3)では、パリティ処理部１９は、ホストから更新ライト要求が発生した際にリードされた、対応するディスク内の更新前データとパリティデータから、更新後のパリティを算出する。そして、世代管理部は、ディスクからリードした更新前データを、そのディスク内のアドレス値と共に、更新前データバッファに格納する。

【００４８】また、世代管理部は、ディスクから更新前データとパリティをリードした後、ディスク媒体が少なくとも１周した際に、ディスク上の同一位置に更新後のデータ及び更新後のパリティを格納し、更新前データバッファがフルになった際にバッファ内のデータをディスクに格納する。

【００４９】このようにすれば、ディスクに格納されているファイルが更新された際に、旧データを効率良く格納し、バックアップやファイルの操作ミスでファイルを破壊した際の復旧のために、旧データを保持し管理するファイルの世代管理を効率良く行える。

【００５０】(d)：前記(4)では、世代管理部の制御手段は、ホストから更新ライト要求が発生した際に、ディスクから対応する更新前データとパリティをリードし、前記パリティ処理部に更新後のパリティを算出させ、ディスクから更新前データとパリティをリードした後、ディスク媒体が少なくとも１周した際に、ディスク上の同一位置に更新後のデータ及びパリティを格納し、ディスクからリードした更新前データを、そのディスク内のアドレス値と共に前記更新前データバッファに格納し、該更新前データバッファがフルになった際に該バッファ内のデータをディスク内の連続領域に格納する。

【００５１】このようにすれば、ファイル内の一部が更新された場合でも、従来のまとめ書き方式（ＷＡＦＬ方式）のように、更新されたデータはディスク内の別領域に格納されているために新たなシークを発生させてしまう、ということがなく、常に、ディスクの連続領域に最新のデータを格納することができる。

【００５２】また、ディスクに格納されているファイルが更新された際に、旧データを効率良く格納し、バックアップやファイルの操作ミスでファイルを破壊した際の復旧のために、旧データを保持し管理するファイルの世代管理を効率良く行える。

【００５３】(e)：以上述べたことから明らかなように、前記(1)、(2)では、ストライプを構成する各ディスクへの格納データの大きさに等しいサイズのまとめ書きバッファのみを用意すれば済むため、まとめ書きバッファ用に用意されるＮＶＲＡＭのような高価な不揮発性メモリの使用量を減らすことができる。また、ディスクへの格納の際に、格納しようとするディスクにリード要求が起こった時に、シークなしでデータを格納できる領域が存在する場合には、その空き領域にデータを格納することによって、ライトの際のシーク時間を短縮することができる。

【００５４】更に、まとめ書きバッファ内で、ホストからの要求アドレス（論理アドレス）によってソート（並べ替え）することによって、ホストにおいて連続領域に格納したデータを、ディスク上のまとめ書き領域においても連続領域に格納することができる。そのため、該当

するデータをリードする際の処理時間を短縮することができる。

【００５５】また、前記(3)、(4)では、前記の点に加え、ファイルの一部が更新された場合でも、常にディスクの連続領域に最新のデータを格納することができ、リード性能が向上する。また、従来のファイル世代管理では、ファイルを新規に作成した時のデータとそのファイルが更新される度に、どこが更新されたかを管理していたが、本願発明では、このような従来の世代管理とは異なり、現ファイルとの差分データのみをディスクへ格納するので、更新前のデータを復元することが可能になる。従って、何世代前までの復元を行う制御が容易にできる。

【００５６】

【発明の実施の形態】以下、発明の実施の形態を図面に基づいて詳細に説明する。

【００５７】§１：システム構成とディスクアレイ装置の説明・・・図２、３参照

図２はシステム構成とディスクアレイ装置の説明図であり、Ａ図はシステム構成図、Ｂ図はディスクアレイ装置のブロック図である。図３は図２の一部詳細図であり、Ａ図はアドレス変換部の説明図、Ｂ図はまとめ書きバッファの説明図である。

【００５８】図２のＡ図に示したシステムは、クライアント・サーバシステムの例であり、ＬＡＮ上に、多数のクライアント装置１６－１、１６－２・・・と、１台のサーバ１５が接続されている。そして、サーバ１５にはディスクアレイ装置７が接続されている。この場合、サーバ１５は、ＬＡＮを介して複数のクライアント装置１６－１、１６－２・・・で使用されている個々のファイルをまとめてディスクアレイ装置７に保持し、各クライアント装置１６－１、１６－２・・・からファイルへのリード／ライト要求を受け付ける、所謂ネットワークサーバを構成している。

【００５９】例えば、前記サーバ１５が、ＵＮＩＸをＯＳ（システムソフトウェア）としたワークステーション（ＷＳ）の場合、ＮＦＳ（Network File System）と呼ばれるプロトコルを用いて各クライアント装置１６－１、１６－２・・・からファイルへのリード／ライト要求をネットワーク（この場合はＬＡＮ）で転送する場合、複数の比較的小さな（８ＫＢや３２ＫＢ程度）データに分割してやりとりする。そのため、サーバ１５がディスクアレイ装置７のディスクへリード／ライト要求を出す場合の多くは、８ＫＢや３２ＫＢ程度のデータ単位で行われる。

【００６０】前記ディスクアレイ装置７の構成を図２のＢ図に示す。前記ディスクアレイ装置７はサーバ１５に接続されて運用されるが、この場合、ディスクアレイ装置７から見ると、前記サーバ１５がホスト（上位装置）となる。前記ディスクアレイ装置７は、ＲＡＩＤ４又はＲＡＩＤ５構成のディスクアレイを記憶手段として保持する装置であり、ディスクアレイ制御装置２と、複数のディスク装置６を備え、前記ディスクアレイ制御装置２は、ディスクアレイコントローラ４と、複数のデバイスアダプタ（ＤＡ）５を備えている。

【００６１】また、ディスクアレイコントローラ４は、パリティ処理を行うためのパリティ処理部１９と、パリティデータを格納するパリティバッファ２３と、論理アドレス（ホスト１から与えられるアドレス）と物理アドレス（実際にディスクに書き込むアドレス）との変換処理を行うアドレス変換部２０と、ホスト１から転送されたデータを一時格納する第１のまとめ書きバッファ２１、及び第２のまとめ書きバッファ２２と、前記第１、第２のまとめ書きバッファ２１、２２内のデータに対する論理アドレスをキーにしてソート（並べ替え処理）を行う論理アドレスソート部１８を備えている。

【００６２】すなわち、サーバ１５は、ＲＡＩＤ４、又はＲＡＩＤ５構成のディスクアレイを記憶手段として保持し、第１のまとめ書きバッファ２１、及び第２のまとめ書きバッファ２２はダブルバッファ構成をとり、片方のまとめ書きバッファがディスクへのデータライトを行っている間にサーバ１５へ到達するライトデータは他方のまとめ書きバッファへ蓄えられる。

【００６３】例えば、第１のまとめ書きバッファ２１のデータをディスクへ書き込んでいる時、いずれかのクライアント装置からのライト要求がサーバ１５に到達した場合、そのライト要求のデータは、第２のまとめ書きバッファ２２に蓄えられる。また、第２のまとめ書きバッファ２２のデータをディスクへ書き込んでいる時、いずれかのクライアント装置からのライト要求がサーバ１５に到達した場合、そのライト要求のデータは、第１のまとめ書きバッファ２１に蓄えられる。

【００６４】前記パリティバッファ２３には、パリティ処理部１９による処理でＲＡＩＤの各ストライプ単位にパリティデータを蓄積する。また、アドレス変換部２０は図３のＡ図に示すように、アドレス変換テーブルを備えており、該アドレス変換テーブルには、サーバ１５のシステムソフトウェア（ＯＳ）が設定したライトデータをディスクへ格納するアドレス（論理アドレス）を、まとめ書きで実際に格納するディスク内のアドレス（物理アドレス）へ変換する情報（論理アドレスと物理アドレスを対応させた情報）が書き込まれるようになっている。

【００６５】また、図３のＢ図に示すように、第１、第２のまとめ書きバッファ２１、２２には、サーバ１５から転送されたライトデータｄ１、ｄ２、ｄ３・・・を格納する。そして、パリティ処理部１９は、これらのライトデータについてのパリティを算出する。例えば、ライトデータｄ１、ｄ２、ｄ３、ｄ４のパリティは、これらのストライプ毎のデータの排他的論理和を算出すること

で求めることができる。
【００６６】また、前記のようにして求めたパリティデータは、パリティ処理部１９によりパリティバッファ２３に格納される。この場合、全てのストライプデータがまとめ書きバッファ内に書き込まれるまでは、前記パリティデータをパリティバッファ２３内に保持する。更に、前記論理アドレスソート部１８では、まとめ書きバッファ内のデータがフル（満杯）になった時点でディスクへの書き込みを行うが、この時、まとめ書きバッファ内のデータを、その論理アドレスをキーにしてソートした後、ディスクへの書き込みを行うようにする。
【００６７】§２：処理概要の説明
以下、ディスクアレイ装置の処理概要を説明する。なお、以下の説明では、サーバ１５を「ホスト」とも記す。ホストからライト要求が発せられると、パリティ処理部１９によりパリティ計算し、ライトデータは、まとめて、第１のまとめ書きバッファ２１、又は第２のまとめ書きバッファ２２内に蓄積され、パリティデータはパリティバッファ２３に格納する。
【００６８】この場合、ホストは前記ライトデータの格納アドレスを指定するが、このホストが指定したアドレス（論理アドレス）は、ディスクアクセスに対する論理的なアドレスとして、アドレス変換部２０のアドレス変換テーブルにセットされ、ディスク内で実際に格納される物理アドレスを割り当てる（アドレス変換テーブルに、論理アドレスと物理アドレスを対応させて格納する）。
【００６９】また、ライトデータは、物理アドレスで指定されるディスク内の位置に格納されるが、まとめ書きバッファの大きさは、ＲＡＩＤを構成する各ディスクに格納するデータサイズになっている。ホストから送られてきたデータは、まとめ書きバッファ内に順に格納されるが、その際、パリティ用バッファ２３内の対応するストライプに格納されたパリティデータと排他的論理和演算を実行し、パリティデータを更新する。但し、まとめ書きバッファがストライプの最初のデータを蓄積する場合には、排他的論理和演算は行わず、ライトデータをパリティ用バッファ２３にも格納する。
【００７０】そして、まとめ書きバッファがフルになった時点で、該まとめ書きバッファ内のデータは、論理アドレスソート部１８により、その論理アドレスをキーにしてソートされた後、１台のディスクに格納される。ディスクへの格納は次のようにする。すなわち、タブルバッファを構成する他方のまとめ書きバッファにホストから転送されたデータが格納され始め、そのまとめ書きバッファがフルになるまでの間に、格納しようとするディスクにリード要求が発生した際、同一シリンダ（シークを発生させずに格納できる範囲）に空き領域があるかどうかを判断し、空き領域がある場合には、その部分に格納する。
【００７１】しかし、そのような空き領域が発見できずに他方のまとめ書きバッファがフルになってしまった場合には、ディスク内の空いている領域にまとめ書きバッファ内のデータを全て格納し、ホストからの新たなライト要求のためのまとめ書きバッファを開放する。また、ストライプを構成する全てのディスクにデータを格納した際には、パリティ用バッファをディスクに格納する。
【００７２】以上述べたように、ストライプ（この例では、ｄ１、ｄ２、ｄ３、ｄ４で１ストライプ）を構成する各ディスクへの格納データの大きさに等しいサイズのまとめ書きバッファのみを用意すれば済むため、まとめ書きバッファ用に用意される不揮発性メモリ（この例では、ＮＶＲＡＭ）の量を減らすことができる。また、ディスクへの格納の際に、格納しようとするディスクにリード要求が起こった時に、シークなしでデータを格納できる領域が存在すれば、その空き領域にデータを格納することによって、ライトの際のシーク時間を短縮することができる。
【００７３】更に、まとめ書きバッファ内で、ホストからの要求アドレス（論理アドレス）によってソート（並べ替え）することによって、ホストにおいて連続領域に格納したデータを、ディスク上のまとめ書き領域においても連続領域に格納することができる。そのため、該当するデータをリードする際の処理時間を短縮することができる。
【００７４】§３：詳細な処理の説明
図４は処理説明図である。以下、図２、図３を参照しながら図４に基づいて詳細な処理を説明する。ネットワーク（この例では、ＬＡＮ）から到着するＮＦＳのライトデータは以下のように処理される。
【００７５】ネットワークからサーバ１５へ送られたライトデータは、更新、新規に関わらず到着順にパリティ処理部１９に入力される。ここでパリティ処理部１９では、ライトデータが格納される予定のストライプについて計算されているパリティデータをパリティバッファ２３から読み出し、ライトデータとの排他的論理和演算を実行してパリティを更新し、パリティバッファ２３内の元のパリティ記憶領域に格納する。
【００７６】前記ライトデータは、まとめ書きバッファ内の物理アドレスを付与された後、更新、新規に関わらず到着順にまとめ書きバッファに格納される。まとめ書きバッファのサイズは、ＲＡＩＤを構成する１つのディスクの連続領域に格納される大きさになっており、サーバ１５のシステムソフトウェア（ＯＳ）が設定した格納アドレス（論理アドレス）はまとめ書きバッファに格納されたアドレス（物理アドレス）と共に、アドレス変換テーブルにセットされる。
【００７７】そして、まとめ書きバッファがライトデータでフル（満杯）になった時点で、まとめ書きバッファへのデータ格納が、ダブルバッファを構成する他方のま

とめ書きバッファに対して開始される。そして、ライトデータでフルになったまとめ書きバッファは、論理アドレスソート部１８により、格納されたライトデータを各々の論理アドレス値によってソートする。その後、ディスクに格納する。

【００７８】例えば、最初に、第１のまとめ書きバッファ２１にライトデータを格納した場合、該第１のまとめ書きバッファ２１がライトデータでフルになると、この第１のまとめ書きバッファ２１に格納されたライトデータを各々の論理アドレス値によってソートした、ディスクに格納する。この時、ホストから転送されたライトデータは、第２のまとめ書きバッファ２２に蓄積される。

【００７９】各クライアント装置からのファイルへのライト要求は、連続した領域への複数のＮＦＳのライト要求としてサーバ１５へ転送される。しかし、サーバ１５では、複数のクライアント装置からの要求を受け付けるため、他のクライアント装置からの要求が割り込む可能性があり、連続したライト要求が必ずしも連続領域へのアクセスとは限らないためである。

【００８０】更に、ディスクへの書き込みは、同一ディスクへのリード要求が発生した際に、近傍（シーク動作を伴わない範囲、例えば同一シリンダ内）に、まとめ書きバッファ内のデータを格納できる連続領域が発見された場合に、その領域に格納する。このようにすることによって、ディスクへのデータ格納の際に必要なシークに要する時間を省略することができる。

【００８１】この際、アドレス変換テーブル内の物理アドレスをディスク内のアドレスに変更する。ダブルバッファを構成する他のまとめ書きバッファがライトデータでフルになってしまう迄に、このような領域が発見できない場合には、シークを行って連続領域にまとめ書きバッファ内のデータを格納する。パリティバッファ２３内のデータも同様にしてディスクに格納する。

【００８２】なお、まとめ書きバッファに格納されたデータは、他のデータのリードに際してリードデータが格納されているディスク内の近傍に空き領域がある場合には、その領域に格納される。そのため、１つのストライプを構成するデータの格納位置がディスク間で異なることがあるが、論理的なストライプが実際に各々のディスクにおいて、どこに格納されるかを示す対応テーブルを設け、そのテーブルを参照することで対処することができる。

【００８３】図４の例では、ネットワークからのライト要求時のライトデータｄ１、ｄ２、ｄ３・・・は、それぞれｎＫＢのブロックデータとして転送されるが、例えば、ｄ１のｎＫＢのライトデータはクライアント装置１６－１からのデータ、ｄ２のｎＫＢのライトデータはクライアント装置１６－２からのデータ、ｄ３のｎＫＢのライトデータはクライアント装置１６－３からのデータである。

【００８４】また、この例ではデータ用ディスクを４台とした例であり、ｄ１、ｄ２、ｄ３、ｄ４からなるストライプのデータを１台のディスク（例えば、１ライト単位＝２～３トラック分）に格納するようになっている。また、前記ｎＫＢは、例えば、ディスクの１２８セクタ分のデータであり、このｎＫＢのデータを単位にして論理アドレスが指定されるようになっている。

【００８５】なお、前記の例では、まとめ書きバッファはダブル構成のバッファで説明したが、３個以上のまとめ書きバッファを使用しても、実施可能である。但し、この場合には、高価なＮＶＲＡＭ等のバッファの数が多くなり、その分、コストアップにつながるので、２個のまとめ書きバッファを使用するのが最適である。

【００８６】（他の例の説明）

§１：他の例の装置の説明・・・図５参照

図５は他の例の装置ブロック図である。この例は、ＲＡＩＤ４、又はＲＡＩＤ５のディスクアレイ装置において、世代管理を行う例であり、図５のように構成する。

【００８７】ディスクアレイ装置７は、ＲＡＩＤ４、或いはＲＡＩＤ５構成のディスクアレイを記憶手段として保持する装置であり、ディスクアレイ制御装置２と、複数のディスク装置６を備え、前記ディスクアレイ制御装置２は、ディスクアレイコントローラ４と、複数のデバイスアダプタ（ＤＡ）５を備えている。また、ディスクアレイコントローラ４は、パリティ処理を行うためのパリティ処理部１９と、パリティデータを格納するパリティバッファ２３と、世代管理部２５と、更新前データバッフィ２６等を備えている。

【００８８】前記世代管理部２５は、データの世代管理を行うものであり、パリティ処理部１９に指示を出してパリティ演算を行わせたり、或いは、ホストから更新データを受け取った際、更新前のデータを更新前データバッファ２６へ格納したりする。なお、他の構成は図２に示したディスクアレイ装置と同じである。

【００８９】§２：処理の説明

世代管理部２５は、ホストからの更新ライト要求が発生した時に、更新前のデータとパリティをディスク装置６のディスクから読み出し、パリティ処理部１９に更新後のパリティを計算させた後、更新前のデータを更新前データバッファ２６に格納し、更新後のデータとパリティをディスクに格納する。

【００９０】更新前データバッファ２６には、更新前のデータと、そのデータが格納されていたディスク上のアドレスに対応させたアドレスをセットすることにより、一定期間毎に更新されたデータの１世代前のデータ（ディスク上のアドレスを含む）を保持する。そして、ホストが或るファイルの１世代前のデータを参照したい場合には、そのファイルに属するデータがセットされているアドレス値を基に、更新前データが格納されている領域を探索する。

【００９１】ホストからの更新データが世代管理部２５に到達すると、世代管理部２５は、更新前のデータとパリティをディスクから読み出す。そして、パリティ処理部１９において、新しいパリティを更新前のデータ、パリティ及び更新後のデータから作成し、パリティバッファ２３に格納する。

【００９２】その後、世代管理部２５は、更新前データを読み出し、ディスク媒体が少なくとも１周した際、同一位置に更新後のデータを格納する。同様にして、更新後のパリティも、ディスク媒体が少なくとも１周する間に、パリティ処理部１９に更新後のパリティを計算させ、同一位置に格納する。但し、更新前データのリードが遅くなった場合には、ディスク媒体が１周以上回転することもあり得る。

【００９３】更新前データは、そのデータを格納していたアドレス値と共に、更新前データバッファ２６へ格納する。更新前データバッファ２６内のデータは、一定期間毎にホストからの命令によって、ディスク内へ格納する。

【００９４】以上のようにすれば、ファイルの一部が更新された場合でも、常にディスクの連続領域に最新のデータを格納することができ、リード性能が向上する。また、従来のファイル世代管理のように、ファイルを新規に作成した時のデータと、そのファイルが更新される度に、どこが更新されたかを順次記憶するのではなく、現ファイルとの差分データのみを格納することによって、更新前のデータを復元することができ、何世代前までの復元を行う制御が容易にできる。

【００９５】

【発明の効果】以上説明したように、本発明によれば次のような効果がある。

【００９６】(1)：請求項１では、第１のまとめ書きバッファと、第２のまとめ書きバッファと、パリティ処理部と、パリティバッファと、アドレス変換テーブルと、論理アドレスソート部とを備え、ホストからライト要求が発せられると、パリティ処理部によりパリティ計算し、ライトデータはまとめて、第１のまとめ書きバッファ、又は第２のまとめ書きバッファ内に蓄積され、パリティデータはパリティバッファに格納する。

【００９７】この時、アドレス変換部は、ホストが指定した論理アドレスと、そのアドレスに対応した物理アドレスとをアドレス変換テーブルにセットする。そして、前記いずれか一方のまとめ書きバッファがフルになったら、該まとめ書きバッファのデータをディスクへ格納するが、この時、該フルになったまとめ書きバッファのデータを、論理アドレスソート部によりソートしてから格納する。

【００９８】このようにすれば、ＲＡＩＤ４、ＲＡＩＤ５のディスクアレイにおいて、書き込み時のオーバーヘッドやリード性能などを改善することが可能になる。

【００９９】(2)：請求項２では、第１の制御手段は、ホストからのライトデータを、いずれか一方のまとめ書きバッファに、新規、更新に関わらず到着順に格納する。そして、第２の制御手段は、前記まとめ書きバッファがフルになった時刻以降に、まとめ書きバッファ内のデータを格納するディスクと同一ディスクに対してリード要求が発生した際、該当するリードデータが格納されているディスク上の位置の近傍に、前記まとめ書きバッファ内の全データを格納する空き領域があるかどうかを判断し、前記空き領域があれば、その空き領域にシーク無しで、前記フルになったまとめ書きバッファのライトデータを格納する。

【０１００】このようにすれば、ＲＡＩＤ４、ＲＡＩＤ５のディスクアレイにおいて、書き込み時のオーバーヘッドやリード性能などを改善することが可能になる。

【０１０１】(3)：請求項３では、パリティ処理部は、ホストから更新ライト要求が発生した際にリードされた、対応するディスク内の更新前データとパリティデータから、更新後のパリティを算出する。そして、世代管理部は、ディスクからリードした更新前データを、そのディスク内のアドレス値と共に更新前データバッファに格納する。

【０１０２】また、世代管理部は、ディスクから更新前データとパリティをリードした後、ディスク媒体が少なくとも１周した際に、ディスク上の同一位置に更新後のデータ及び更新後のパリティを格納して、更新前データバッファがフルになった際にバッファ内のデータをディスクに格納する。

【０１０３】このようにすれば、ディスクに格納されているファイルが更新された際に、旧データを効率良く格納し、バックアップやファイルの操作ミスでファイルを破壊した際の復旧のために、旧データを保持し管理するファイルの世代管理を効率良く行える。

【０１０４】(4)：請求項４では、世代管理部の制御手段は、ホストから更新ライト要求が発生した際に、ディスクから対応する更新前データとパリティをリードし、更新後のパリティを算出して、ディスクから更新前データとパリティをリードした後、ディスク媒体が少なくとも１周した際に、ディスク上の同一位置に更新後のデータ及びパリティを格納し、ディスクからリードした更新前データを、そのディスク内のアドレス値と共に、前記更新前データバッファに格納し、該更新前データバッファがフルになった際に、該バッファ内のデータをディスク内の連続領域に格納する。

【０１０５】このようにすれば、ファイル内の一部が更新された場合でも、従来のまとめ書き方式（ＷＡＦＬ方式）のように、更新されたデータはディスク内の別領域に格納されているために新たなシークを発生させてしまう、ということがなく、常に、ディスクの連続領域に最新のデータを格納することができる。

【０１０６】また、ディスクに格納されているファイルが更新された際に、旧データを効率良く格納し、バックアップやファイルの操作ミスでファイルを破壊した際の復旧のために、旧データを保持し管理するファイルの世代管理を効率良く行える。

【０１０７】(5)：請求項１、２では、更に、ストライプを構成する各ディスクへの格納データの大きさに等しいサイズのまとめ書きバッファのみを用意すれば済むため、まとめ書きバッファ用に用意されるＮＶＲＡＭのような高価な不揮発性メモリの使用量を減らすことができる。

【０１０８】また、ディスクへの格納の際に、格納しようとするディスクにリード要求が起こった時に、シークなしでデータを格納できる領域が存在する場合には、その空き領域にデータを格納することによって、ライトの際のシーク時間を短縮することができる。

【０１０９】更に、まとめ書きバッファ内で、ホストからの要求アドレス（論理アドレス）によってソート（並べ替え）することによって、ホストにおいて連続領域に格納したデータを、ディスク上のまとめ書き領域においても連続領域に格納することができる。そのため、該当するデータをリードする際の処理時間を短縮することができる。

【０１１０】(6)：請求項３、４では、更に、ファイルの一部が更新された場合でも、常にディスクの連続領域に最新のデータを格納することができ、リード性能が向上する。また、従来のファイル世代管理では、ファイルを新規に作成した時のデータとそのファイルが更新される度に、どこが更新されたかを管理していたが、本願発明では、このような従来の世代管理とは異なり、現ファイルとの差分データのみをディスクへ格納するので、更新前のデータを復元することが可能になる。従って、何世代前までの復元を行う制御が容易にできる。

【０１１１】(7)：ＲＡＩＤ４、ＲＡＩＤ５のディスクアレイにおいて、書き込み時のオーバーヘッドやリード性能などを改善することができる。また、ディスクに格納されているファイルが更新された際に、旧データを効率良く格納し、バックアップやファイルの操作ミスでファイルを破壊した際の復旧のために、旧データを保持し管理するファイルの世代管理を効率良く行える。

【図面の簡単な説明】

【図１】本発明の原理説明図である。

【図２】本発明の実施の形態におけるシステム構成とディスクアレイ装置の説明図である。

【図３】図２の一部詳細図である。

【図４】本発明の実施の形態における処理説明図である。

【図５】本発明の他の例の装置ブロック図である。

【図６】従来のディスクアレイ装置の説明図である。

【図７】ＲＡＩＤのレベル説明図である。

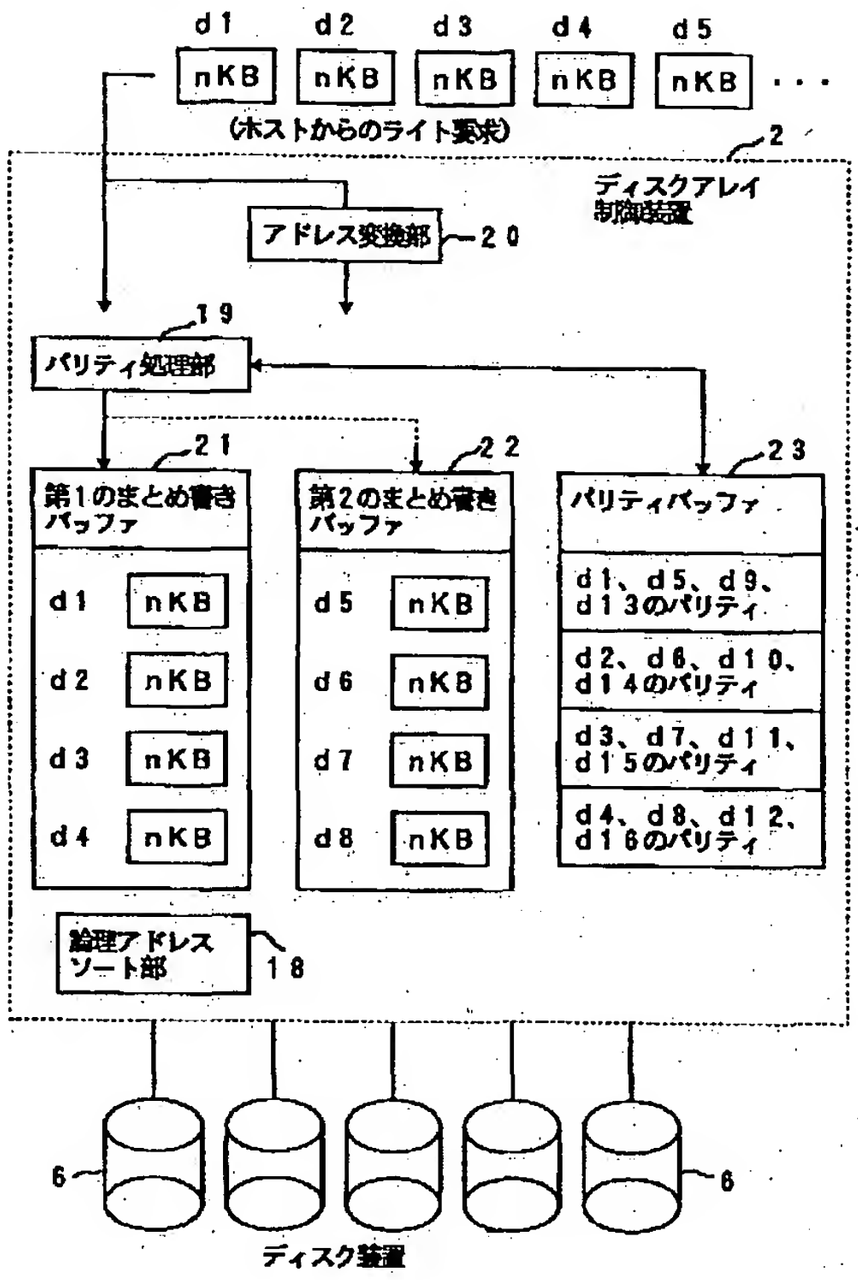
【図８】ＷＡＦＬの説明図である。

【符号の説明】

１　ホスト（ホストコンピュータ）
２　ディスクアレイ制御装置
３　ホストアダプタ
４　ディスクアレイコントローラ
５、５－１～５－ｎ　デバイスアダプタ
６、６－１～６－ｎ　ディスク装置
７　ディスクアレイ装置
８　データバッファ
９　アドレス変換テーブル
１０　まとめ書きバッファ
１５　サーバ
１６　クライアント装置
１８　論理アドレスソート部
１９　パリティ処理部
２０　アドレス変換部
２１　第１のまとめ書きバッファ
２２　第２のまとめ書きバッファ

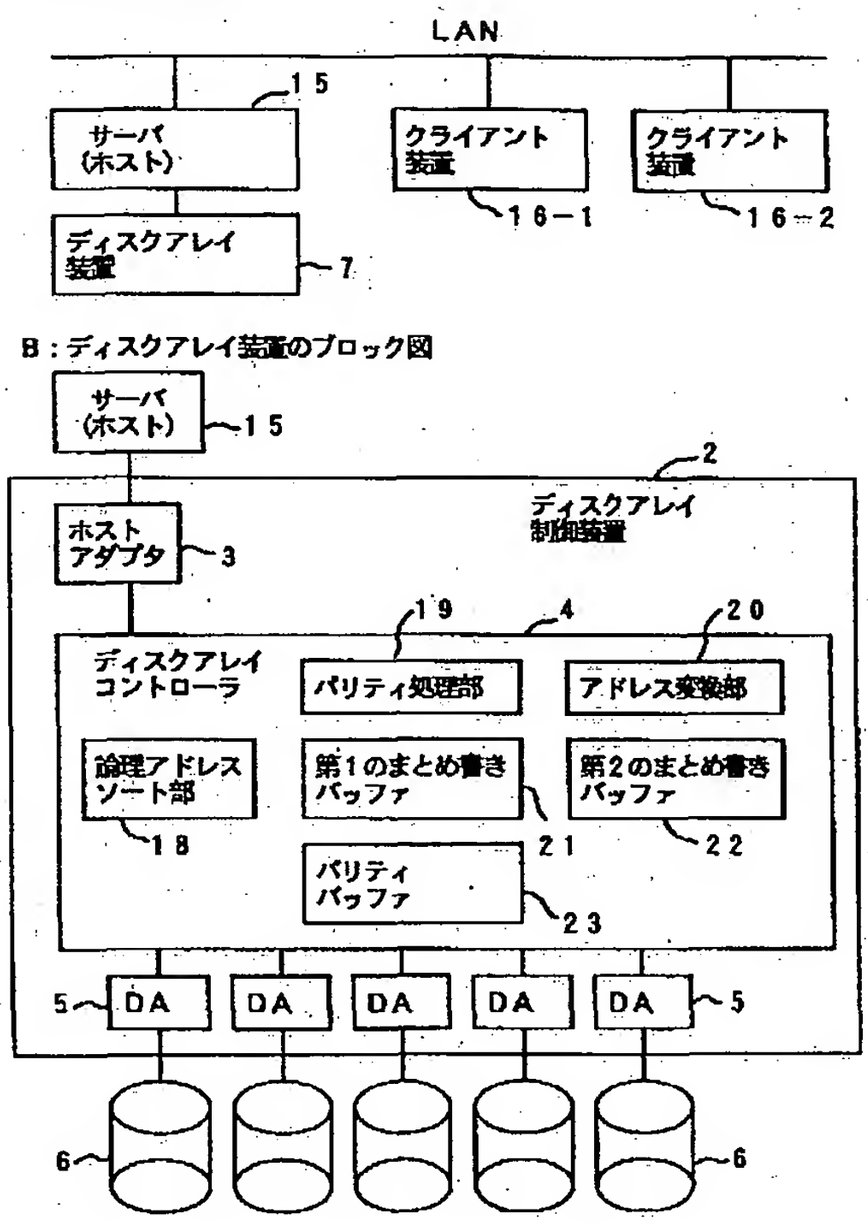
【図１】

本発明の原理説明図



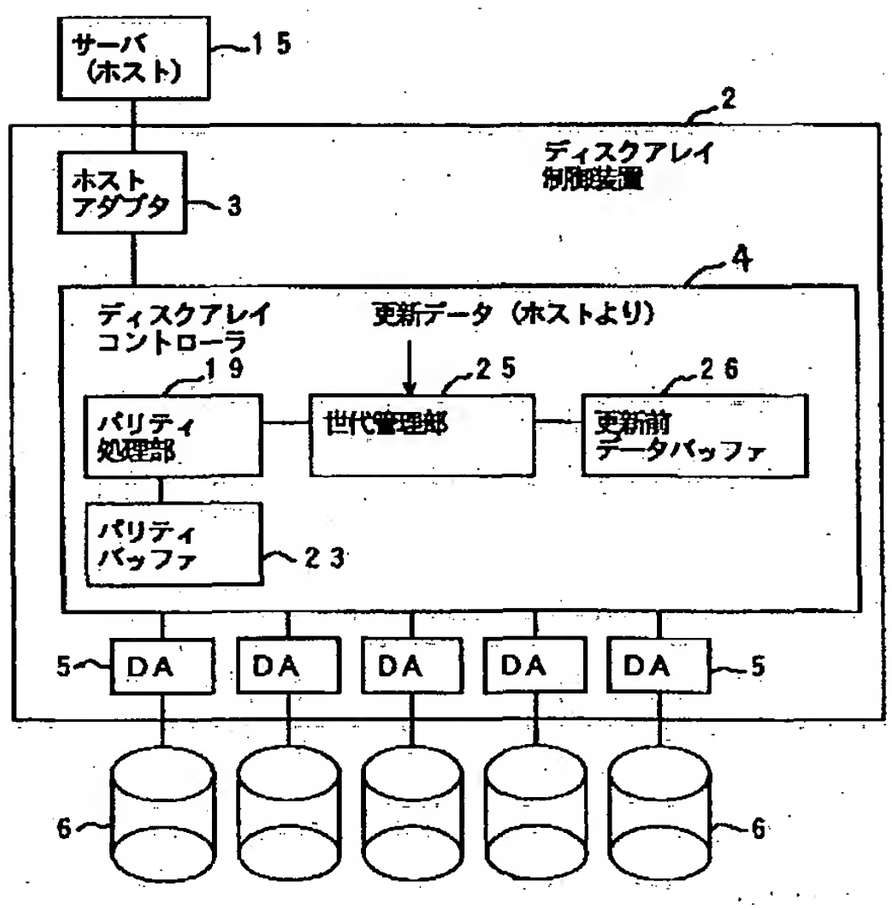
【図２】

システム構成とディスクアレイ装置の説明図



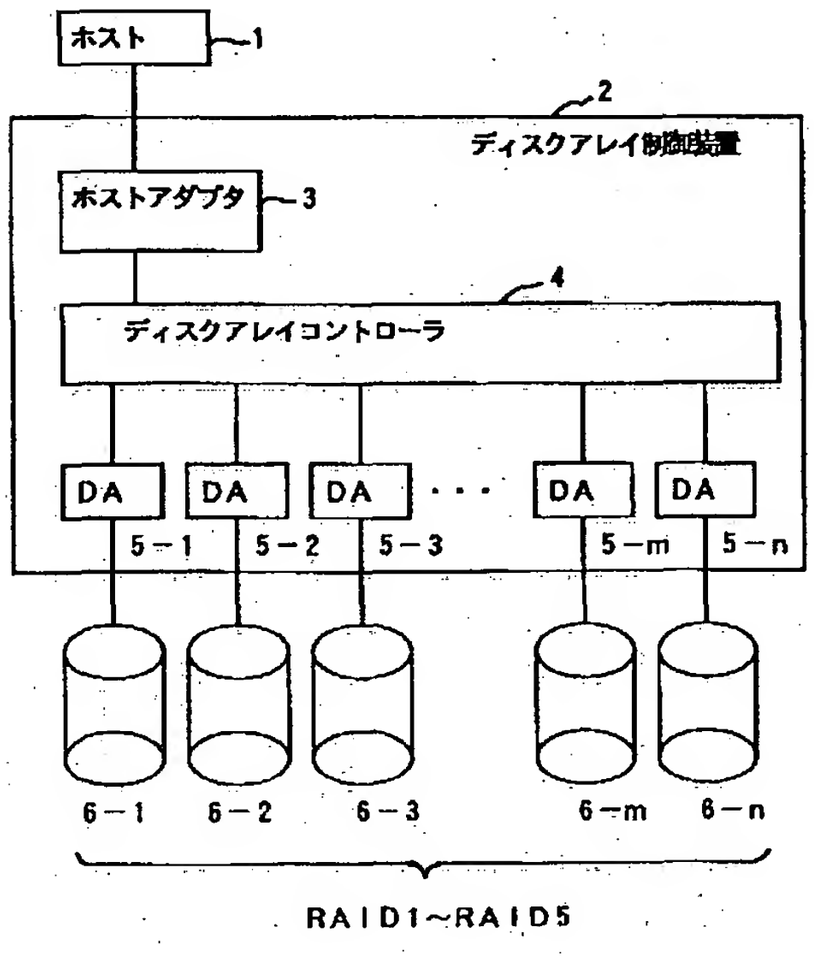
【図５】

他の例の装置ブロック図



【図６】

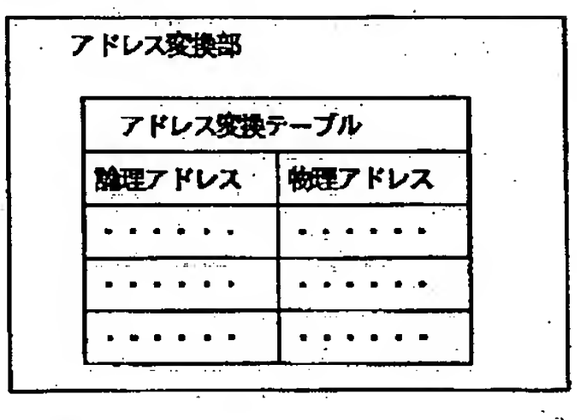
従来のディスクアレイ装置の説明図



【図３】

図2の一部詳細図

A：アドレス変換部の説明図



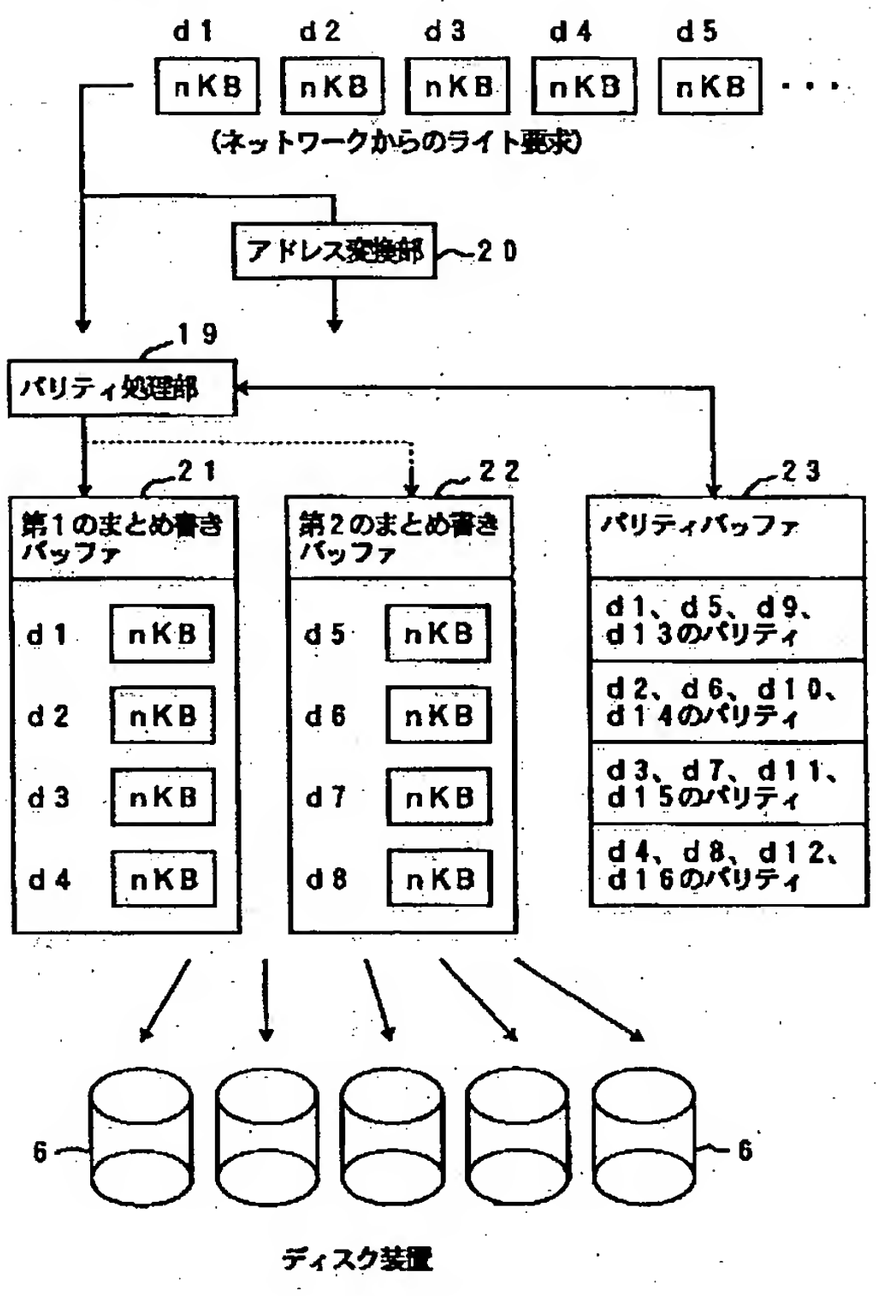
B：まとめ書きバッファの説明図

まとめ書きバッファ				
ライトデータ				パリティ
d1	d2	d3	d4	d1、d2、d3、d4のパリティ
d5	d6	d7	d8	d5、d6、d7、d8のパリティ

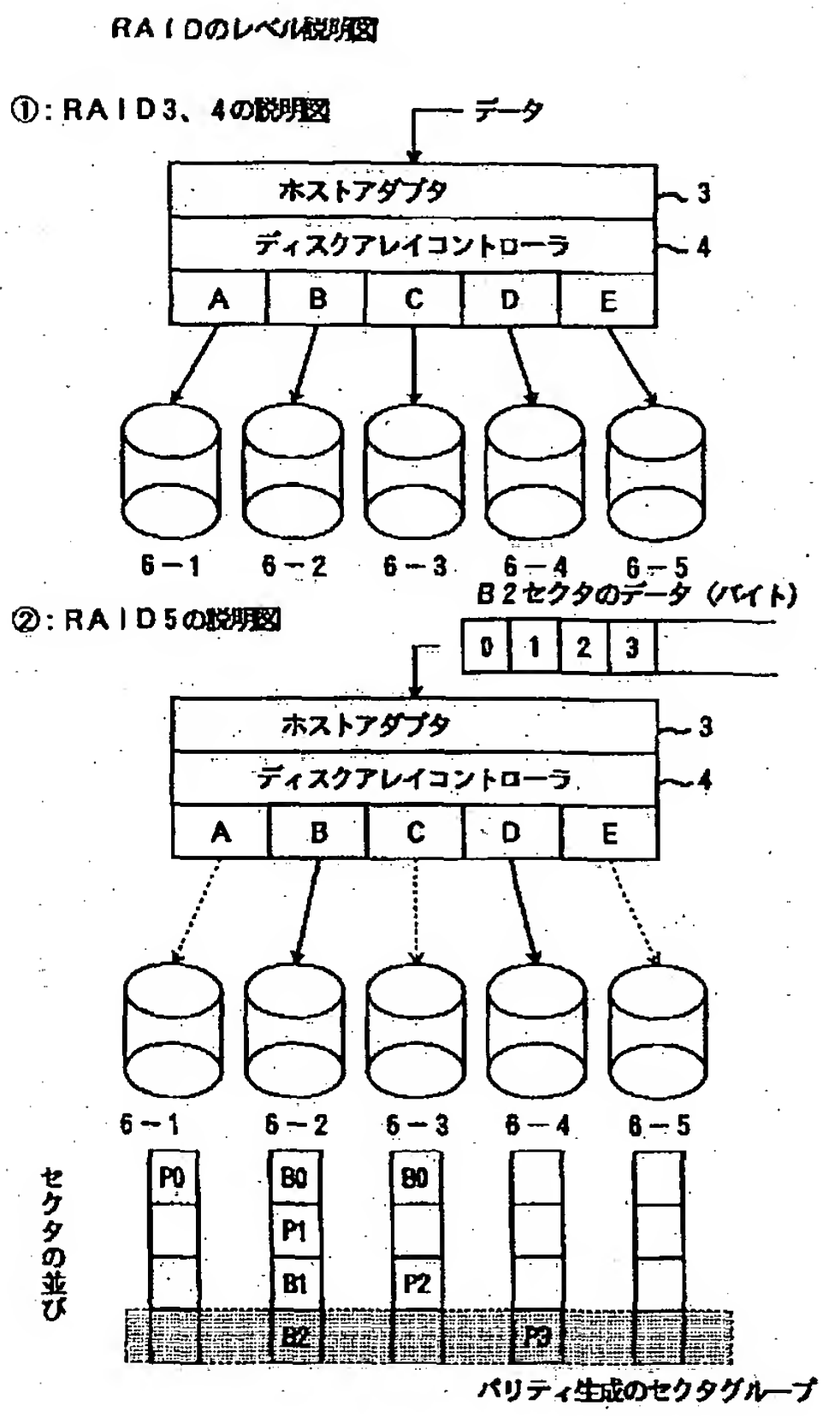
d1～d8：ライトデータ

【図４】

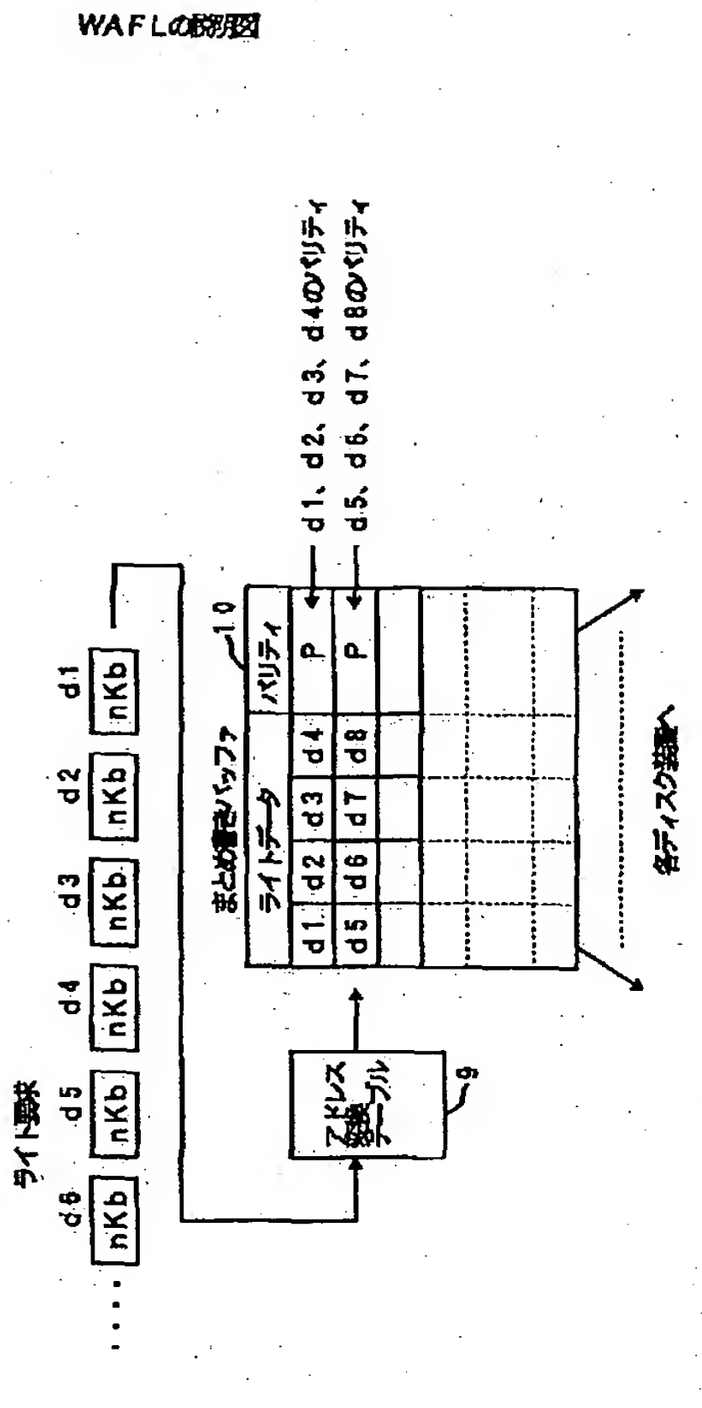
処理説明図



【図７】



【図８】



フロントページの続き

(72)発明者　青木　隆浩
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

Ｆターム(参考)　5B065 BA01 CA30 CC02 CC08 CE16
CH15 EA02 ZA17